# Nonlinear global Fréchet regression for random objects via weak conditional expectation

Satarupa Bhattacharjee, Bing Li, and Lingzhou Xue

Department of Statistics, The Pennsylvania State University
University Park, PA 16802, U.S.A.

**Abstract**

We introduce a nonlinear global regression model for object-valued predictor and response tuples. Random object data are complex non-Euclidean data taking value in general metric space, possibly devoid of any underlying vector space structure. Such data are getting increasingly abundant with the rapid advancement in technology. Examples include probability distributions, positive semi-definite matrices, and data on Riemannian manifolds. We propose the notion of a weak conditional Fréchet mean to aid the object regression framework. One of the main contributions is to establish a connection between the conditional Fréchet mean and the weak conditional Fréchet mean, the latter can being a generalization of the former. The motivation is based on Carleman operators and their inducing functions in the particular case of the classical Euclidean data. The state-of-the-art global Fréchet regression approach by Petersen and Müller (2019) emerges as a special case of the proposed model. We require that the metric space where the predictors reside admits a reproducing kernel Hilbert space embedding that is rich enough to characterize the joint probability distribution of the responses and the predictors, while the intrinsic geometry of the metric space where the responses lie is utilized to study the asymptotic convergence of the proposed estimates. Numerical studies, including both simulations and a data application, are conducted to investigate the performance of our estimator in a finite sample.

## 1 Introduction

Encountering complex non-Euclidean data-taking values in a general metric space that may defy any inherent linear structure has become increasingly common in the areas such as biological or social sciences. With the rapid advancement of technology, we encounter an abundance of such complex random objects recorded as shapes, time-courses of images, and networks. Examples of such "random object" data (Marron and Alonso, 2014) include distributional data in Wasserstein space (Delicado and Vieu, 2017; Le Gouic and Loubes, 2017), symmetric positive definite matrix objects (Dryden et al., 2009), data on the surface of the sphere (Di Marzio et al., 2014), phylogenetic trees (Billera et al., 2001), and finite-dimensional Riemannian manifolds objects (Afsari, 2011), among others. Since the data are metric space valued, many classical notions of statistics, such as the definition of sample or population mean as an average or expected value, do not apply anymore and need to be replaced by barycenters or Fréchet means (Fréchet, 1948). Similarly, for random object responses $Y$ residing in a metric space $(\Omega_Y, d_Y)$ and Euclidean predictors $X \in \mathbb{R}^p$, their intrinsic regression relationship is quantified by modeling the conditional Fréchet means (Hein, 2009; Petersen and Müller, 2019) as

$$m_\oplus(x) = \mathrm{argmin}_{y\in\Omega_Y} E[d_Y^2(Y, y)|X = x];\ x \in \mathbb{R}^p. \tag{1}$$

The Fréchet regression proposed by Petersen and Müller (2019) generalizes the global least squares and the nonparametric local linear regression to fit the conditional Fréchet mean. The globally linear approach targets an alternative formulation than (1) given by

$$\tilde{m}_\oplus(x) = \mathrm{argmin}_{y\in\Omega_Y} E[s(X, x)d_Y^2(Y, y)], \tag{2}$$

where the weight function $s(X, x) = 1+(x-\mu_X)^\intercal\Sigma_X^{-1}(X-\mu_X)$ varies globally and linearly with the output points $x \in \mathbb{R}^p$; $\mu_X$ and $\Sigma_X$ being the expectation and covariance matrix for the predictors $X$.

Model (2) coincides with model (1) in the special case of multiple linear regression with Euclidean responses and predictors. However, for a general metric space-valued response $Y \in \Omega_Y$, the above two targets are different, thus making the regression relationship for general metric-valued data quite restrictive. Although the local regression, which indeed targets (1) with an asymptotically negligible bias, is more flexible, it is effective only when the dimension of the predictor is relatively low. As this dimension gets higher, its accuracy drops significantly- a phenomenon known as the curse of dimensionality. Recently Bhattacharjee and Müller (2021) developed a single index Fréchet regression that projects the multivariate predictors onto a desired direction parameter vector to form a single index, thus facilitating inference for Fréchet regression. However, the model assumptions are still somewhat restrictive, and in general, the Fréchet regression framework only can accommodate Euclidean predictors.

In this work, we propose a non-linear global object regression framework that can accommodate both responses and predictors residing in arbitrary metric spaces. Our main two contributions are listed as follows.

Firstly, as discussed before, the conditional Fréchet mean in (1) might be a significantly different target from the global Fréchet mean (2) proposed by Petersen and Müller (2019), owing to the lack of linearity in a general abstract metric space. Hence the interpretation or validity of such a "globally linear" model can be brought into question. We propose a significant step up in bridging the discrepancy between two targets and extending the global linear regression to a more general globally non-linear object regression.

In order to answer this question, one first needs to ponder what a polynomial regression model even looks like in a metric space. A convenient vehicle to link random object data analysis to non-linear global RKHS (Reproducing Kernel Hilbert Space) regression models, beyond linear or polynomial regression to an arbitrary non-linear function, is achieved through weak conditional moments on $d_Y^2(Y, \omega)$. Li and Song (2022) first introduced this new statistical construct as a generalization of conditional expectation based on Carleman operators and their induced functions. The key idea of this approach for classical Euclidean data is that it replaces the $L_2$ space for the projection that characterizes the conditional expectation by an arbitrary Hilbert space, while still maintaining the unbiasedness of the regression estimate. We define a random object regression model by extending this concept to the Weak conditional Fréchet mean via a kernelized version of the predictors. The global linear regression model by Petersen and Müller (2019) emerges as a special case of a linear kernel.

Secondly, beyond scalar-or-vector-valued predictors, studying the relation between two arbitrary random objects is also increasingly important. Unfortunately, not much exists in the state-of-the-art literature in this context, barring special cases of distribution-on-distribution regression (Chen et al., 2019, 2021; Ghodrati and Panaretos, 2022). Our proposed method accommodates more general predictors such as random vectors, functions, or even object-valued predictors, as long as the predictor space is rich enough to admit an RKHS embedding. We discuss the details of constructing appropriate kernels to generate such RKHSs and study the relevant operators generated to achieve this goal.

The rest of the paper is organized as follows. Section 2 defines the preliminary setup of the problem and focuses on the construction of an RKHS on the space predictor objects. In Section 3, we define the weak condition moments for object responses and predictors, establish the global non-linear object regression model and study its connections to the global linear object regression framework. In Section 4, we propose a suitable estimator for the weak conditional Fréchet mean from the observed data. In this vein, the construction of the underlying RKHS is discussed and an M-estimation setting is devised. Section 5 establishes the asymptotic convergence rates of the proposed methods. Simulation results are presented in Section 6 to show the numerical performances of the proposed methods. In Section 7, we analyze a real application of the proposed method for the mortality-vs-fertility distributions. All proofs are presented in the Supplementary Material.

# 2 Preliminaries

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $(\Omega_X, \mathcal{F}_X)$ and $(\Omega_Y, \mathcal{F}_Y)$ be two measurable metric spaces. Define $X : \Omega \to (\Omega_X, d_X)$ and $Y : \Omega \to (\Omega_Y, d_Y)$ to be the random objects. We denote the marginal distributions of $X$ and $Y$ by $P_X$ and $P_Y$, respectively, the joint distributions of $(X, Y)$ by $P_{XY}$, and the conditional distributions of $Y|X$ by $P_{Y|X}$. Further, let $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$ be a positive definite kernel and $\mathcal{H}_X$ the Reproducing Kernel Hilbert Space (RKHS) of real-valued functions on the predictor space $(\Omega_X, d_X)$ generated by $\kappa_X$. We aim for direct modeling of the joint distribution $P_{XY}$ by introducing the concept of the weak conditional Fréchet regression, which involves the intrinsic geometry of $(\Omega_Y, d_Y)$ as well as the appropriate operators defined on the RKHS $\mathcal{H}_X$. In the next subsection, we will describe the construction of such RKHS.

## 2.1 Random operators and their moments

A natural way to construct a reproducing kernel on $(\Omega_X, d_X)$ is to take a classical radial basis function, such as the Gaussian radial basis kernel $\kappa_X(s_1, s_2) = e^{-\gamma||s_1 - s_2||}$, and replace the Euclidean distance therein with the distance $d_X$ in the metric space. Metric spaces that are of negative type can produce positive definite kernels of such form. For example, Theorem 4 of Kolouri et al. (2016) states that Wasserstein space of absolutely continuous univariate distributions can be isometrically embedded in a Hilbert space, and thus the Gaussian RBF kernel constructed on such space is positive definite. For commonly observed random object data such as distributional data, the construction of such kernels is well studied (Zhang et al., 2022).

Further, the RKHS $\mathcal{H}_X$ generated by universal kernels, such as the Gaussian radial basis function, is shown to be dense in $L_2(P_X)$ (Sriperumbudur et al., 2010), where $L_2(P_X)$ be the class of square-integrable functions of $X$ under the measure $P_X$. More generally, Proposition 2 of Zhang et al. (2021) shows that if $(\Omega_X, d_X)$ is a complete separable metric space and there exists a separable Hilbert space$\mathcal{H}$ and a continuous injection $\rho : \Omega_X \to \mathcal{H}$, then then the kernel function $\kappa_X : \Omega_X \times \Omega_X : \mathbb{R}$ defined by $\kappa_X(s_1, s_2) = F(\langle \rho(s_1), \rho(s_2) \rangle_{\mathcal{H}})$ is a positive definite kernel, where $F$ is an analytic function whose Taylor series at zero has strictly positive coefficients. Let $\kappa_G(x, x') = \exp(-\gamma_X d_X^2(x, x'))$ and $\kappa_L(x, x') = \exp(-\gamma_X d_X^2(x, x'))$ denote the Gaussian and Laplacian kernels, respectively. Zhang et al. (2021) showed that both $\kappa_G$ and $\kappa_L$ on a complete and separable metric space $\Omega_X$ are positive definite and universal, and the RKHS $\mathcal{H}_X$ generated by such kernels is dense in $L_2(P_X)$.

In order to capture the nonlinear features of a random element, we define the covariance operator in the RKHS $\mathcal{H}_X$, which is similar to the construction in Fukumizu et al. (2004); Lee et al. (2013); Li and Song (2017); Sang and Li (2022). See Li (2018) Chapter 12.2 for a comprehensive review. For two generic Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_1$, let $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ denote the class of bounded linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$. If $\mathcal{H}_1 = \mathcal{H}_1 = \mathcal{H}$, we denote $\mathcal{B}(\mathcal{H}, \mathcal{H})$ as $\mathcal{B}(\mathcal{H})$. For any operator $T \in \mathcal{B}(\mathcal{H})$, let $Y^*$ denote the adjoint operator of $T$, $\ker(T)$ the kernel of $T$, $\text{ran}(T)$ the range of $T$, and $\overline{\text{ran}}(T)$ the closure of the range of $T$. For two members $f$ and $g$ of $\mathcal{H}$, the tensor product $f \otimes g$ is the operator on $\mathcal{H}$ such that $(f \otimes g)h = f\langle g, h \rangle_{\mathcal{H}}$ for all $h \in \mathcal{H}$.

Further, we define $E(\kappa_X(\cdot, X))$, the expected value of $X$ in $\mathcal{H}$, as the unique element in $\mathcal{H}_X$ given by the Riesz representation theorem, such that

$$\langle f, E(\kappa_X(\cdot, X)) \rangle_{\mathcal{H}_X} = E\left[\langle f, \kappa_X(\cdot, X) \rangle_{\mathcal{H}_X}\right], \text{ for all } f \in \mathcal{H}_X. \tag{3}$$

Define the bounded linear operator $E\left[\kappa_X(\cdot, X) \otimes \kappa_X(\cdot, X)\right]$, the second-moment operator of $X$ in $\mathcal{H}_X$, as the unique element in $\mathcal{B}(\mathcal{H}_X)$ by virtue of Riesz representation theorem such that, for all $f$ and $g$ in $\mathcal{H}_X$,

$$\langle f, E\left[\kappa_X(\cdot, X) \otimes \kappa_X(\cdot, X)\right] \rangle_{\mathcal{H}_X} = E\left[\langle f, (\kappa_X(\cdot, X) \otimes \kappa_X(\cdot, X))g \rangle_{\mathcal{H}_X}\right]. \tag{4}$$

By Cauchy-Schwartz inequality and Jensen's inequality, it is guaranteed that items on the right-hand side of (3) and (4) are well-defined. We denote by $\mu_X = E(\kappa_X(\cdot, X))$, and $M_{XX} = E\left[\kappa_X(\cdot, X) \otimes \kappa_X(\cdot, X)\right]$. The auto-covariance operator is defined as

$$\Sigma_{XX} = M_{XX} - \mu_X \otimes \mu_X = E\left[(\kappa_X(\cdot, X) - E(\kappa_X(\cdot, X))) \otimes (\kappa_X(\cdot, X) - E(\kappa_X(\cdot, X)))\right]. \tag{5}$$

Using this, for all $f, g \in \mathcal{H}_X$, we have $\mathrm{cov}(f(X), g(X)) = \langle f, \Sigma_{XX} g\rangle_{\mathcal{H}_X}$.

The auto-covariance operator in (5) will be revisited while defining the weak conditional Fréchet regression in Section 3. Before proceeding further, we briefly review the notion of weak conditional moments in the context of random functions as Hilbert space objects in the next subsection. This notion will then be generalized as weak conditional Fréchet means for random metric-space valued objects in the later sections.

## 2.2 A brief review of weak conditional moments

For any Hilbertian objects $U$ and $V$, the weak conditional expectation is defined as the inducing function of a Carleman operator (Weidmann [36]). In the literature on sufficient dimension reduction, the key advantage of this approach is that it replaces the $L_2$ -space for the projection that characterizes the conditional expectation by an arbitrary Hilbert space while still maintaining the unbiasedness of the dimension reduction estimate. Intuitively, the method proposed in Li and Song (2022) is reiterated as follows. A feature space based on the random function $V$ using a reproducing kernel is constructed, and the projection of the random function $U$ is carried out onto the feature space through the tensor product $U \otimes U$. The projections reduce to conditional expectations if the feature space is sufficiently large, but if not, they still provide valid estimates of the dimension reduction space.

To define more formally, let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $U : \Omega \to \Omega_U$ and $V : \Omega \to \Omega_V$ be Borel random functions. Further, let $\kappa_V : \Omega_V \times \Omega_V \to \mathbb{R}$ be a positive definite kernel function and let $\mathcal{H}_V$ be the reproducing kernel Hilbert space generated by the kernel $\kappa_V$. Further, let $\mathcal{H}_U$ be a Hilbert space and $U \in \mathcal{H}_U$. Let $M_{VU} = E[\kappa_V(\cdot, v) \otimes U]$ and $M_{VV} = E[\kappa_V(, V) \otimes \kappa_V(\cdot, V)]$. Under the assumption that $\ker(M_{VV}) = \{0\}$ and $\mathrm{ran}(M_{VU}) \subset \mathrm{ran}(M_{VV})$, $M_{VV} : \mathcal{H}_V \to \mathrm{ran}(M_{VV})$ is a one-to-one transformation. The first condition in the above assumptions is satisfied if $\kappa_V$ is a continuous kernel, while the second condition requires that, for any $f \in \mathcal{H}_U$,

$M_{VU}f$ is sufficiently concentrated on the low- frequency components of $M_{VV}$, imposing smoothness in the space. We define the Moore-Penrose inverse map $M_{VV}^{\dagger}$ to be the linear operator from $\mathrm{ran}(M_{VV})$ to $\mathcal{H}_V$ such that, for any $f \in \mathrm{ran}(M_{VV})$, $M_{VV}^{\dagger}$ is the unique $g \in \mathcal{H}_V$ satisfying $f = M_{VV}g$. Under the condition, $\mathrm{ran}(M_{VU}) \subset \mathrm{dom}(M_{VV}^{\dagger})$, the operator $R_{VU} = M_{VV}^{\dagger}M_{VU}$ is well defined. This operator is often referred to as the regression opertor (Lee et al., 2016).

**Definition 1 (Carleman ocperator)** *Let $\mathcal{G}$ be a set, $\mathcal{M}$ a Hilbert space of real-valued functions on $\mathcal{G}$, and $A : \mathcal{H} \to \mathcal{M}$ a linear operator for some other Hilbert space $\mathcal{H}$. If, for each $y \in \mathcal{G}$, the linear functional*

$$A_y : \mathcal{H} \to \mathbb{R}, \ f \mapsto (Af)(y)$$

*is bounded, then we call $A$ an extended Carleman operator. The Riesz representation $\lambda_A(y)$ of $A_y$ is called the inducting function of $A$.*

The above definition is slightly more general since any subset of the class of square-integrable functions on $\mathcal{G}$ is not required to be a subspace of $L_2(P_V)$, the class of square-integrable functions of $V$ under its marginal measure $P_V$. In this context, such a subset of functions is taken as the RKHS $\mathcal{H}_V$, whose inner product is different from that of $L_2(P_V)$.

**Definition 2 (Weak conditional moments)** *If the regression operator $R_{VU}$ is a Carleman operator, then the random element $V \mapsto \lambda_{R_{VU}}(V)$, the inducing function of the Carleman operator $R_{VU}$ is called the weak conditional moment of $U$ given $V$ and is written as $E[U ¦ V]$.*

By Weidmann (2012), Theorem 6.12, $M_{VV}^{\dagger}A$ is a Carleman operator if it is a Hilbert Schmidt operator, for any operator $A$ that maps into the domain of $M_{VV}^{\dagger}$, which is a reasonable assumption, as this amounts to imposing a type of smoothness in the relation between $U$ and $V$ (Li and Song, 2017). In connection with the regression operator, $R_{VU}$, Li and Song (2022) and some of the references therein showed the following result.

**Lemma 1** *If $f \in \mathcal{H}_V$, $E(f(V)|U = \cdot)$ is a member of $\mathcal{H}_U$, and $\kappa_U$ is a universal kernel such that $\kappa_U : \Omega_U \times \Omega_U \to \mathbb{R}$, then*

$$R_{VU}f = E[f(V)|U = \cdot] + \textit{constant}.$$

By taking unconditional expectation on both sides of the above equality, we see that the constant is $E[(R_{VU}f)(U)] - E[f(V)]$, and we have the formula

$$E[f(V)|U = \cdot] = R_{VU}f - E[(R_{VU}f)(U)] + E[f(V)].$$

When $\kappa_U$ is not universal, $R_{VU}f - E[(R_{VU}f)(U)] + E[f(V)]$ is not the conditional expectation. Nevertheless, Li and Song (2022) shows that it shares many properties with the conditional expectation, particularly those pertaining to the regression on $V$ on $U$. Li and Song (2022) call $R_{VU}f - E[(R_{VU}f)(U)] + E[f(V)]$ the weak conditional expectation of $f(V)$ relative to $\mathcal{H}_X$, and denote it by $E[f(V)\text{¦}U]$.

# 3 Weak conditional Fréchet mean

The above section is only discussed for a comprehensive understanding of weak conditional moments in the special case of Hilbert space-valued data. More generally, we want to draw motivation from this concept to extend it for metric-space valued object data $(X, Y) \in (\Omega_X, d_X) \times (\Omega_Y, d_Y)$, as defined in Section 2, with RKHS $(\mathcal{H}_X)$ embedding in the predictor space $\Omega_X$ via the positive definite kernel $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$.

The flexibility of the above construct is particularly important for such random object data, since attempting to estimate the conditional Fréchet mean function $E[d_Y^2(Y, \omega)|X]$ is often inefficient without assuming a structural form due to the curse of dimensionality. The weak conditional expectation gives us an idea of performing regression of $Y$ on $X$ when $Y$ is a random object.

**Definition 3** *Let $h : \Omega_X \to \Omega_Y$ be defined by*

$$h(x) = \text{argmin}_{y\in\Omega_Y} E[d_Y^2(Y, \omega)\text{¦}X = x]$$

*is called the weak Fréchet conditional expectation of $Y$ given $X$, and denote it by $E^{(F)}(Y\text{¦}X)$.*

We take the weak Fréchet conditional expectation as the target of estimation in our Fréchet regression. Let us now derive a more explicit form of $E^{(F)}(d_Y^2(Y, \omega)\text{¦}X)$.

**Theorem 1** *Suppose $(\Omega_Y, d_Y)$ is a metric space, $\mathcal{H}_X$ is an RKHS on $\Omega_X$ generated by a positive kernel $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$, and $\mathcal{H}_Z$ is the RKHS generated by the linear kernel $\kappa_Z(z_1, z_2) = c + z_1^\intercal z_2$, where $Z = d_Y^2(Y, \omega)$. Then, for any $y \in \Omega_Y$,*

$$E[d_Y^2(Y, \omega)\text{¦}X = x] = E[d_Y^2(Y, \omega)] + \langle \kappa_X(\cdot, x) - \mu_X, \Sigma_{XX}^\dagger E[(\kappa_X(\cdot, X) - \mu_X))\, d_Y^2(Y, \omega)]\rangle_{\mathcal{H}_X} \tag{6}$$

*Consequently, if $E[d_Y^2(Y,\omega)|X=\cdot]$ is a member of $\mathcal{H}_X$ and $\kappa_X$ is a universal kernel, then*

$$E[d_Y^2(Y,\omega)|X=x] = E[d_Y^2(Y,\omega)] + \langle \kappa_X(\cdot,x)-\mu_X, \Sigma_{XX}^{\dagger} E[(\kappa_X(\cdot,X)-\mu_X)\, d_Y^2(Y,\omega)]\rangle_{\mathcal{H}_X}.$$

Define the following operators to quantify the interaction between the operator of $X$ in $\mathcal{H}_X$ and the positive real-valued distance elements $d_Y^2(Y,\omega)$ and $d_Y(Y,\omega)$ as

$$\Sigma_{XY}^{(k)}(\omega) = E\left[(\kappa_X(\cdot,X)-\mu_X)\, d_Y^k(Y,\omega)\right], \text{ for all } \omega\in\Omega_Y;\ k=1,2. \tag{7}$$

Even though $d^k(Y,\omega)$ do not represent measures of similarity as a kernel would do in the context, as such $\Sigma_{XY}(\omega)$ are not exactly the same as cross-covariance operators. However, these operators quantify the association between the (mean-subtracted) kernel embedding, which can be perceived as similarity measures in the predictor space, and the distance between the responses and any other element in the corresponding metric space. Our aim is to search over the latter metric space to find the minimizer of the weighted expected distance in view of the weak conditional Fréchet expectation defined in Theorem 1. In that sense, the aim is to still minimize a suitable version of the expected error in $Y$, which is not explained by the predictors $X$. Henceforth, we will call the above quantities pseudo-cross covariance operators, which will be useful in the subsequent lemmas and theorems in the next sections.

Typically, for a positive definite $\kappa_X$, $\Sigma_{XX}$ is a compact operator whose eigenvalues decay to 0, hence $\Sigma_{XX}^{\dagger}$ is unbounded. However, it is reasonable to assume that the regression operators $R_{XY}^{(1)}(\omega) := \Sigma_{XX}^{\dagger}\Sigma_{XY}^{(1)}(\omega)$ and $R_{XY}^{(2)}(\omega) := \Sigma_{XX}^{\dagger}\Sigma_{XY}^{(2)}(\omega)$ to be bounded uniformly for all $\omega\in\Omega_Y$. These can be viewed as regression operators, due to their similarity in appearance to the coefficients arising in multiple linear regression.

Reiterating from Theorem 1, the weak conditional Fréchet mean in (6) can be rewritten in terms of the auto covariance and pseudo-cross covariance operators defined in Section 2 as

$$E[d_Y^2(Y,\omega)] + \langle \kappa_X(\cdot,x)-\mu_X, \Sigma_{XX}^{\dagger}\Sigma_{XY}^{(2)}(\omega)\rangle_{\mathcal{H}_X}.$$

Suppose the eigenvalue and eigenfunction sequence of $\Sigma_{XX}$ is given by $\{(\lambda_j,\phi_j):\ j=1,2,\ldots\}$. By Mercer's theorem, the spectral decomposition of the variance operator $\Sigma_{XX}$ is given by

$$\Sigma_{XX} = \sum_{j=1}^{\infty} \lambda_j \phi_j \otimes \phi_j. \tag{8}$$

Further, under the assumption that $E(\kappa_X(X,X)) < \infty$ (see Assumption (A0) below), $\Sigma_{XX}$ is a Hilbert-Schmidth operator and is of trace class, i.e, $\sum_{j=1}^{\infty} \lambda_j < \infty$.

(A0) $E(\kappa_X(X,X)) < \infty$ and $\sup\limits_{\omega \in \Omega_Y} E(d_Y^k(Y,\omega)) < \infty$ for $k = 1, 2$.

We further assume a degree of smoothness in the relation between $X$ and $Y$, requiring the output functions for the regression operator must be sufficiently concentrated on the low-frequency components of $\Sigma_{XX}$. We assume that

(A1) $\sup\limits_{\omega \in \Omega_Y} E\left(|\phi_j(X) - E(\phi_j(X))|\ d_Y^k(Y,\omega)\right) \leq \lambda_j^2$, $k = 1, 2$.

i.e., $R_{XY}^{(k)}(\omega) := \Sigma_{XX}^{\dagger}\Sigma_{XY}^{(k)}(\omega)$; $k = 1, 2$ is a bounded operator uniformly for all $\omega \in (\Omega_Y, d_Y)$, in other words $\operatorname{ran}(\Sigma_{XY}^{(k)}(\omega))$, which can possibly depend on $\omega$ is entirely contained in the $\operatorname{ran}(\Sigma_{XX})$ uniformly across all possible $\omega \in \Omega_Y$, for $k = 1, 2$. Even though $\Sigma_{XX}^{\dagger}$ can be an unbounded operator., it never appears by itself but is always accompanied by operators multiplied from the right to appear as a regression operator. Condition (A1) guarantees that that the composite operators $\Sigma_{XX}^{\dagger}\Sigma_{XY}^{(k)}(\omega)$ is well-defined, bounded and compact, uniformly for all $\omega \in \Omega_Y$, for $k = 1, 2$. This implies that $\Sigma_{XX}^{\dagger}\Sigma_{XY}^{(k)}(\omega)$ must send all incoming functions into the low-frequency range of the eigenspaces of $\Sigma_{XX}$ with relatively large eigenvalues uniformly for all $\omega \in \Omega_Y$, for $k = 1, 2$. That is, $\Sigma_{XY}^{(k)}(\omega)$ is smooth uniformly for all $\omega \in \Omega_Y$ in the sense that its outputs are low-frequency components of $\Sigma_{XX}$, for $k = 1, 2$.

In the special case where $\kappa_X$ is the linear kernel $c + x_1^{\intercal}x_2$, $E^{(F)}(d_Y^2(Y,\omega) \mid X)$ reduces to the objective function for global linear Fréchet regression. To prove this result, we first introduce an isomorphism between $\operatorname{ran}(\Sigma_{XX})$ and $\mathbb{R}^p$, where $\Sigma_{XX}$ and $\mathcal{H}_X$ are defined by the linear kernel. Su et al. (2023) proved the following result.

**Proposition 1** *Let $\mathcal{H}_X$ be the RKHS generated by $\kappa_X(x_1, x_2) = c + x_1^{\intercal}x_2$. Let $\Sigma_{XX} : \mathcal{H}_X \to \mathcal{H}_X$ be the covariance operator of $X$, and let $\Sigma_X$ be the matrix* $\operatorname{var}(X)$. *Let $\mathcal{H}_0 = \operatorname{ran}(\Sigma_{XX})$. Then*

1. *$\mathcal{H}_0 = \operatorname{span}\{(\cdot)^{\intercal}y : y \in \mathbb{R}^p\}$.*

2. *Let $T : \mathcal{H}_0 \to \mathbb{R}^p$ be the mapping $(\cdot)^{\intercal}y \mapsto y$. Then $T$ is an isomorphism.*

3. *$T\Sigma_{XX}T^* = \Sigma_X$.*

The following proposition shows how to invert the covariance operator $\Sigma_{XX}$ in the linear kernel setting.

**Proposition 2** *If $\Sigma_X$ is an invertible matrix then $\Sigma_{XX}$ is an invertible operator, and*

$$\Sigma_{XX}^{-1} = T^{*}\Sigma_X^{-1}T.$$

The next theorem shows that the global Fréchet linear regression introduced by Petersen and Müller (2019) is, in fact, the weak Fréchet conditional mean in the special case when $\kappa_X$ is taken to be the linear kernel.

**Theorem 2** *Suppose $\Sigma_X$ is invertible, $\kappa_X$ and $\kappa_Y$ are the linear kernels. Then*

$$E[d_Y^2(Y,\omega) \mathrel{¦} X = x] = E\left\{[1 + (x - EX)^{\intercal}\Sigma_X^{-1}(X - EX)]d_Y^2(Y,\omega)\right\}.$$

## 3.1 Existence of the weak Fréchet conditional means

When $\kappa_X$ is any arbitrary kernel such as a linear kernel and is not necessarily a universal kernel, the weak conditional Fréchet mean $h(x) = \operatorname*{argmin}_{y\in\Omega_Y} E(d_Y^2(Y,\omega) \mathrel{¦} X = x)$ is not same as the conditional Fréchet mean $m(x) = \operatorname*{argmin}_{y\in\Omega_Y} E(d_Y^2(Y,\omega)|X = x)$. For example, the target for the global Fréchet regression, which emerges as a special case of the weak conditional Fréchet means corresponding to a linear kernel, is different from the conditional Fréchet regression function $E(d_Y^2(Y,\omega)|X = x)$. However, the regression relationship between two random objects $(X, Y) \in \Omega_X \times \Omega_Y$ expressed through the weak Fréchet conditional mean is interesting and worth investigating in its own right. This alternative formulation is described through an RKHS embedding in the predictor space, thus accommodating random objects lying in the general metric space as a predictor. The characterization of the dependence between $Y$ and $X$ is global and non-linear, and no bandwidth parameter is required to fine-tune the regression function.

The existence of the weak conditional Fréchet mean is guaranteed when the response lies in a compact metric space $\Omega_Y$, as long as the function $E(d_Y^2(Y,\omega) \mathrel{¦} X = x)$ is a continuous function of $y$. Beyond compactness, the existence and uniqueness of the weak conditional Fréchet mean can be proved for commonly observed random objects with the explicit forms of the minimizers available.

**Definition 4 (Negative type metric space)** *The space $(M, \rho)$ with a semi-metric $\rho$ is of negative type if for all $n \geq 2$, $z_1, z_2, \ldots, z_n \in M$ and $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^n \alpha_i = 0$, one has $\sum_{i=1}^n \sum_{j=1}^n \alpha_i\alpha_j\rho(z_i, z_j) \leq 0$.*

**Proposition 3** *If $\Omega_Y$ is a negative type metric space and there is a continuous injective map $\rho : \Omega_Y \to \mathcal{F}$ for some underlying Hilbert space $\mathcal{F}$, such that the image of $\rho$ is a closed and convex set, then the minimizer of the weak conditional Fréchet mean exists and is unique.*

Some examples of commonly observed random object data, where the explicit solutions for the minimizers are available include the following.

*Example 1:* The space of univariate probability distributions $G$ on $\mathbb{R}$ such that $\int_{\mathbb{R}} x^2 G(x) < \infty$, equipped with the Wasserstein-2 metric. For two such distributions $G_1$ and $G_2$, the Wasserstein-2 metric between $G_1$ and $G_2$ is given by

$$d_W^2(G_1, G_2) = \int_0^1 (G_1^{-1}(t) - G_2^{-1}(t))^2 dt, \tag{9}$$

where $G_1^{-1}$ and $G_2^{-1}$ are the quantile functions corresponding to $G_1$ and $G_2$, respectively. The weak conditional Fréchet mean for distributional objects endowed with the Wasserstein-2 metric $d_W$ as defined above is given by

$$\begin{aligned} h(x) =& \underset{Q_\omega \in Q(\Omega_Y)}{\operatorname{argmin}} \; E\left(d_W^2(Q_Y, Q_\omega) \mid X = x\right) \\ =& E(Q_Y(t)) + \langle \kappa_X(\cdot, x) - \mu_X, \; \Sigma_{XX}^{\dagger} \; E\left((\kappa_X(\cdot, X) - \mu_X) Q_Y(t)\right)\rangle_{\mathcal{H}_X}. \end{aligned}$$

*Example 2:* The space of symmetric positive semi-definite matrices $\mathcal{M}$ endowed with the Frobenius metric $d_F$. For any two elements $A, B \in (\mathcal{M}, d_F)$, their Frobenius distance is given by

$$d_F^2(A, B) = \sqrt{\text{trace } ((A - B)(A - B)^T)}. \tag{10}$$

The weak conditional Fréchet mean for spd matrix objects equipped with the Frobenius metric $d_F$ is given by

$$h(x) = \underset{y \in \Omega_Y}{\operatorname{argmin}} \; E\left(d_F^2(Y, \omega) \mid X = x\right),$$

where $h(x)$ has the $(j, k)$-th entry as

$$B_{jk}(x) = E(Y_{jk}) + \langle \kappa_X(\cdot, x) - \mu_X, \; \Sigma_{XX}^{\dagger} \; E\left((\kappa_X(\cdot, X) - \mu_X) Y_{jk}\right)\rangle_{\mathcal{H}_X},$$

where $Y_{jk}$ is the $(j, k)$-th entry of the positive semi-definite matrix response $Y \in (\Omega_Y, d_F)$. The existence, uniqueness, and explicit form of the weak conditional Fréchet mean can also be derived for other Euclidean and pseudo-Euclidean metrics such as power metric, log-affine metric, Cholesky metric and so on (Dryden et al., 2010; Lin, 2019).

When the kernel $\kappa_X$ is universal such as the Gaussian or Laplacian kernels, the RKHS $\mathcal{H}_X$ is rich enough so that the proposed model for the weak conditional Fréchet mean can approximate the conditional Fréchet mean arbitrarily closely (see e.g., Li and Song (2022)) since the RKHS $\mathcal{H}_X$ is dense in $L_2(P_X)$. In the case when the conditional Fréchet mean falls within $\mathcal{H}_X$, the weak conditional Fréchet mean coincides with the conditional Fréchet mean.

# 4 Estimation

In the last section, we have described the solution to the nonlinear object regression framework at the population level. In the following, we implement the regression at the sample level. The key steps involve the construction of the sample estimate for the regression function as an M-estimator based on $n$ i.i.d. observations $(X_i, Y_i)_{i=1}^n$. In order to quantify the sample objective function minimized by the regression estimator, we need to express the underlying RKHS $\mathcal{H}_X$ and the relevant auto covariance and pseudo-cross covariance operators with a coordinate representation system (see, e.g., Horn and Johnson (2012); Li (2018)).

## 4.1 Coordinate representation

Suppose that $\mathcal{L}_1$ is a finite dimensional linear space with basis $\mathcal{B} = \{\xi_1, \xi_2, \ldots, \xi_p\}$. Then for any $\xi \in \mathcal{L}_1$, there is a unique vector $(a_1, a_2, \ldots, a_p)^\intercal \in \mathbb{R}^p$ such that $\xi = \sum_{i=1}^p a_i \xi_i$. The vector $(a_1, a_2, \ldots, a_p)^\intercal$ is called the coordinate of $\xi$ with respect to $\mathcal{B}$, and denoted by $[\xi]_{\mathcal{B}}$. Throughout this section, we will use this notation to describe coordinate representation. Next, we introduce the coordinate representation of a linear operator between two (finite-dimensional) linear spaces. Suppose $\mathcal{L}_2$ is another linear space with basis $\mathcal{C} = \{\eta_1, \eta_2, \ldots, \eta_q\}$ and $A$ is a linear operator from $\mathcal{L}_1$ $\mathcal{L}_2$.

Then for any $\eta \in \mathcal{L}_1$, we have

$$\begin{aligned} A\xi = A\left(\sum_{i=1}^{p} \left([\xi]_{\mathcal{B}}\right)_i \xi_i\right) &= \sum_{i=1}^{p} \left([\xi]_{\mathcal{B}}\right)_i (A\xi_i) \\ = \sum_{i=1}^{p} \left([\xi]_{\mathcal{B}}\right)_i \sum_{j=1}^{q} \left([A\xi_i]_{\mathcal{C}}\right)_j \eta_j &= \sum_{j=1}^{q} \left\{ ({}_{\mathcal{C}}[A]_{\mathcal{B}}) \left([\xi]_{\mathcal{B}}\right) \right\}_j \eta_j, \end{aligned}$$

where ${}_{\mathcal{C}}[A]_{\mathcal{B}}$ is the $q \times p$ matrix with $(i,j)$th entry $([A\xi_j]_{\mathcal{C}})_i$. The above equation implies that $[A\xi]_{\mathcal{C}} = ({}_{\mathcal{C}}[A]_{\mathcal{B}})([\xi]_{\mathcal{B}})$. Therefore we call the matrix ${}_{\mathcal{C}}[A]_{\mathcal{B}}$ the coordinate representation of the linear operator $A$ with respect to the bases $\mathcal{B}$ and $\mathcal{C}$. Similarly, for two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, with spanning systems $\mathcal{B}_1$ and $\mathcal{B}_2$, and a linear operator $A : \mathcal{H}_1 \to \mathcal{H}_2$, we use the notation ${}_{\mathcal{B}_1}[A]_{\mathcal{B}_2}$ to represent the coordinate representation of $A$ relative to spanning systems $\mathcal{B}_1$ and $\mathcal{B}_2$.

## 4.2 Construction of the RKHS $\mathcal{H}_X$ and Model Fitting

Recall that for a positive definite kernel defined on a set $T \times T$ denoted by $\kappa_T$. Let $K_T$ be the $m \times m$ Gram matrix whole $(k,l)$-th entry is $\kappa_T(\nu_k, \nu_l)$ for $\nu_k, \nu_l \in T$. Let $H_T$ be the RKHS generated by $\{\kappa_T(\cdot, \nu_k) : k = 1, \ldots, m\}$. Then using the coordinate representation in Section 4.1, the inner product between any $f, g \in \mathcal{H}_T$ can be expressed as $\langle f, g\rangle_{\mathcal{H}_T} = [f]^\intercal K_T [g]$.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. observations of $(X, Y) \in \Omega_X \times \Omega_Y$. The RKHS $\mathcal{H}_X$ is spanned by $\{\kappa_X(\cdot, X_i) : i = 1, \ldots .n\}$ equipped with the inner product

$$\langle f, g\rangle_{\mathcal{H}_X} = [f]^\intercal K_X [g],$$

for any $f, g \in \mathcal{H}_X$, where $K_X$ is the $n \times n$ Gram matrix whose $(i,j)$th entry is $\kappa_X(X_i, X_j)$, $i, j = 1, \ldots, n$.

At the sample level, we estimate $\Sigma_{XX}$ and $\Sigma_{XY}(\omega)$ by replacing the expectations $E(\cdot)$ with the sample moments $E_n(\cdot)$ with respect to the empirical measure whenever possible. For example, we estimate $\Sigma_{XX}$ by $\hat{\Sigma}_{XX} = E_n\left[(\kappa_X(\cdot, X) - \hat{\mu}_X) \otimes (\kappa_X(\cdot, X) - \hat{\mu}_X)\right]$ $= \frac{1}{n}\sum_{i=1}^{n} (\kappa_X(\cdot, X_i) - \hat{\mu}_X) \otimes (\kappa_X(\cdot, X_i) - \hat{\mu}_X)$, where $\hat{\mu}_X = E_n(\kappa_X(\cdot, X_i)) = \frac{1}{n}\sum_{i=1}^{n} \kappa_X(\cdot, X_i)$. The sample estimate for $\Sigma_{XY}^{(k)}(\omega)$ for any given $\omega \in \Omega_Y$ is similarly defined as $\hat{\Sigma}_{XY}^{(k)}(\omega) =$ $E_n\left[(\kappa_X(\cdot, X) - \hat{\mu}_X) d^k(Y, \omega)\right] = \frac{1}{n}\sum_{i=1}^{n} (\kappa_X(\cdot, X_i) - \hat{\mu}_X) d^k(Y_i, \omega)$, for $k = 1, 2$. Suppose, the subspace $\overline{\mathrm{ran}}(\hat{\Sigma}_{XX})$ is spanned by the set $\mathcal{B}_X = \{\kappa_X(\cdot, X_i) - E_n(\kappa_X(\cdot, X_i)) : i = 1, \ldots, n\}$.

We then have the following coordinate representations of auto covariance and pseudo-cross covariance operators, for any $\omega \in \Omega_Y$ and for $k = 1, 2$,

$$ {}_{\mathcal{B}_X}[\hat{\Sigma}_{XX}]_{\mathcal{B}_X} = n^{-1} G_X,\ [\hat{\Sigma}^{(k)}_{YX}(\omega)]_{\mathcal{B}_X} = n^{-1} G_X,\ {}_{\mathcal{B}_X}[\hat{\Sigma}^{\dagger}_{XX}]_{\mathcal{B}_X} = n^{-1} G^{\dagger}_X, $$

where $G_X = QK_XQ$ and $G^{\dagger}_X$ is the Moore-Penrose inverse of $G_X$ via the Tikhonov-regularized inverse $(G_X + \epsilon_X I_n)^{-1}$ to prevent overfitting, where $\epsilon_X > 0$ is a tuning constant. Here $Q$ denotes the projection matrix $I_n - \frac{1}{n} 1_n 1_n^T$ with $Q^2 = Q$. For a detailed discussion see e.g. Section 12.4 of Li (2018).

Now, we proceed to estimate the weak conditional Fréchet mean in (6). Recalling the definition from Section 3,

$$ \begin{aligned} h(x) &= \underset{\omega \in \Omega_Y}{\operatorname{argmin}}\ J(\omega),\ \text{where} \\ J(\omega) &= E[d_Y^2(Y, \omega)] + \langle \kappa_X(\cdot, x) - \mu_X, \Sigma^{\dagger}_{XX} E[(\kappa_X(\cdot, X) - \mu_X)\, d_Y^2(Y, \omega)] \rangle_{\mathcal{H}_X}; \end{aligned} \tag{11} $$

we define the estimator

$$ \begin{aligned} \hat{h}(x) &= \underset{\omega \in \Omega_Y}{\operatorname{argmin}}\ J_n(\omega),\ \text{where} \\ J_n(\omega) &= \frac{1}{n} \sum_{i=1}^{n} d_Y^2(Y_i, \omega) + \langle \kappa_X(\cdot, x) - \hat{\mu}_X, \hat{\Sigma}^{\dagger}_{XX} \frac{1}{n} \sum_{i=1}^{n} (\kappa_X(\cdot, X_i) - \hat{\mu}_X)\, d_Y^2(Y_i, \omega) \rangle_{\mathcal{H}_X} \\ &= \frac{1}{n} \sum_{i=1}^{n} w_{in}(x) d_Y^2(Y_i, \omega), \end{aligned} \tag{12} $$

where $w_{in}(x) = 1 + \langle \kappa_X(\cdot, x) - \hat{\mu}_X, \hat{\Sigma}^{\dagger}_{XX}(\kappa_X(\cdot, X_i) - \hat{\mu}_X) \rangle_{\mathcal{H}_X}$. To obtain a more explicit computable form of the above, it remains to identify the coordinate of $\kappa_X(\cdot, x) - \hat{\mu}_X$ with respect to the spanning system $\{\kappa_X(\cdot, X_i) - \hat{\mu}_X : i = 1, \ldots, n\}$. Suppose that $[\kappa_X(\cdot, x) - \hat{\mu}_X] = c_x$ for some $c_x \in \mathbb{R}^n$. Then

$$ \langle \kappa_X(\cdot, x) - \hat{\mu}_X, \kappa_X(\cdot, X_i) - \hat{\mu}_X \rangle_{\mathcal{H}_X} = e_i^{\intercal} K_X c_x - \frac{1}{n} (e_i^{\intercal} K_X 1_n)(1_n^{\intercal} c_x) = e_i^{\intercal} K_X Q c_x, $$

where $e_i$ denotes the vector whose $i$th component is 1 and all others are 0. Taking $i = 1, \ldots, n$, we have $d_x = K_X Q c_x$, where $d_x$ is the vector of length $n$ with $i$th component $\kappa_X(X_i, x) - E_n(\kappa_X(X_i, x))$. With the Tikhonov regularization, we obtain the solution $c_x = Q(K_X + \epsilon_X I_n)^{-1} d_x$. Thus, the empirical objective function in (12) becomes

$$ J_n(\omega) = \frac{1}{n} h_Y^{\intercal} 1_n + h_Y^{\intercal} G_X (G_X + \epsilon_X I_n)^{-1} c_x, $$

where $h_Y$ is the vector with the $i$-th component $d_Y^2(Y_i, \omega)$, $i = 1, \ldots, n$, and $1_n = (1, 1, \ldots, 1)^{\intercal}$.

## 4.3 Tuning parameter selection

We use the general cross-validation criterion (Golub et al., 1979) to determine the tuning constant $\epsilon_X$ involved in the Tikhonov-regularization of the inverse auto-covariance operator $\Sigma_{XX}^{\dagger}$.

$$\mathrm{GCV}(\epsilon_X) = \frac{1}{n}\sum_{i=1}^{n} \frac{d_Y^2(Y_i, \hat{Y}_i)}{\left(1 - \mathrm{tr}[QG_X(G_X + \epsilon_X I_n)^{-1} + 1_n 1_n^{\intercal}/n]/n\right)^2}, \tag{13}$$

where $Y_i$ and $\hat{Y}_i$ are respectively the observed and predicted responses for the $i$-th subject, $i = 1, \ldots, n$. The numerator of this criterion quantifies the prediction error while the denominator controls the degree of overfitting. We minimize the criterion over a grid $\{10^{-6}, \ldots, 10^{-1}\}$ to find the optimal tuning constants.

# 5 Convergence results

In this section, we develop the asymptotic convergence results for the proposed object regression method. In particular, the convergence of the auto-covariance and pseudo-cross-covariance operators with a suitable rate is established, which is used in turn to show the convergence of the regression estimate using the M-estimation theory.

## 5.1 Convergence of regression operators

Recalling the definitions from section 2.1, $\mu_X = E(\kappa_X(\cdot, X))$ and $\Sigma_{XX} = E\big[\big(\kappa_X(\cdot, X) - E(\kappa_X(\cdot, X))\big) \otimes (\kappa_X(\cdot, X) - E(\kappa_X(\cdot, X)))\big]$ are the mean and auto covariance operator defined on the RKHS $\mathcal{H}_X$. The asymptotic properties of the empirical estimates of the relevant quantities have been well-studied in the literature (see e.g., Sang and Li (2022); Fukumizu et al. (2007); Lee et al. (2013). For completion, we list the properties here

**Lemma 2** *Under Assumptions (A0) and (A1)*

*(1)* $\|\hat{\mu}_X - \mu_X\|_{\mathcal{H}_X} = O_P(n^{-1/2})$.

*(2)* $\|\hat{\Sigma}_{XX} - \Sigma_{XX}\|_{OP} = O_P(n^{-1/2})$.

The consistent estimation for the pseudo-cross covariance operators defined in Section 2.1 is derived uniformly over all elements $\omega \in \Omega_Y$, under the following assumption on the intrinsic geometry and complexity of the response space $(\Omega_Y, d_Y)$, which can be quantified by a bound on the entropy integral of $\Omega_Y$.

(K0) The entropy integral of $\Omega_Y$ is finite, i.e.,

$$J := \int_0^1 \sqrt{1 + \log N(\epsilon, \Omega_Y, d)} d\epsilon < \infty,$$

where $N(\epsilon, \Omega_Y, d)$ is the covering number for the space $\Omega_Y$ using balls of radius $\epsilon$.

This assumption is satisfied by most of the commonly observed random objects such as the space of univariate distributions with Wasserstein metric, space of positive semi-definite matrices with a suitable choice of metric, data on the surface of an $n-$sphere with the intrinsic geodesic metric, and so on (see e.g. Dubey and Müller (2019) and the references therein).

**Lemma 3** *Under Assumptions (A0), (A1), and (K0),*

$$\sup_{\omega \in \Omega_Y} \|\hat{\Sigma}_{XY}^{(k)}(\omega) - \Sigma_{XY}^{(k)}(\omega)\|_{OP} = O_P(n^{-1/2}),\ \ k = 1, 2.$$

The consistent estimation for the regression operators is described in the following lemma, under sufficient smoothness conditions on the regression relationship between $X$ and $Y$. We assume

(A2) For all $j \in \mathbb{N}$, there is a $0 < \beta \leq 1$ such that $\sup_{\omega \in \Omega_Y} E\left(|\phi_j(X) - E(\phi_j(X))|\, d_Y^{(k)}(Y, \omega)\right)$ $\leq \lambda_j^{2+\beta}$, for $k = 1, 2$, i.e. there is a bounded linear operator $S_{XY} : \mathcal{H}_X \to \mathcal{H}_X$ such that $\sup_{\omega \in \Omega_Y} \Sigma_{XX}^{(1+\beta)^\dagger} \Sigma_{XY}^{(k)}(\omega)$ is a bounded linear operator uniformly over all $\omega \in \Omega_Y$ for $k = 1, 2$.

Suppose $n^{-1/2} \prec \epsilon_n \prec 0$. For any $\beta$ as defined in Assumption (A2), define

$$\alpha_n = \epsilon_n^\beta + \epsilon_n^{-1} n^{-1/2}. \tag{14}$$

**Proposition 4** *Under Assumptions (A0), (A1), (A2), and (K0),*

$$\sup_{\omega \in \Omega_Y} \|\hat{\Sigma}_{XX}^\dagger \hat{\Sigma}_{XY}^{(k)}(\omega) - \Sigma_{XX}^\dagger \tilde{\Sigma}_{XY}^{(k)}(\omega)\|_{OP} = O_P(\alpha_n);\ \ k = 1, 2,$$

*where $\alpha_n$ is as given in* (14).

## 5.2 Estimation of weak conditional Fréchet mean

Having established the convergence of the pseudo-regression operators $\Sigma_{XX}^{\dagger}\Sigma_{XY}^{(k)}$, $k = 0, 1$, we proceed to derive the convergence results for the weak Fréchet conditional mean in (12). We require the following assumptions regarding the intrinsic geometry of the response space, which are the key to establishing the rate of convergence of any M-estimator, namely, the assumption of well-separateness of the minimizer, an upper bound on the entropy integral of the underlying metric space, and a local lower bound on the curvature of the objective functions.

(R1) The weak conditional Fréchet means $h(x)$ and $\hat{h}(x)$ exist and are unique, the latter almost surely. Further, the minimizer at the population level is well separated. i.e., for any $\epsilon > 0$,

$$\inf_{d_Y(\omega, h(x)) > \epsilon} J(\omega, x) - J(h(x), x) > 0.$$

(R2) Let $B_\delta(h(x)) \subset \Omega_Y$ be the ball of radius $\delta$, centered at $h(x)$ and $N(\epsilon, B_\delta(h(x)), d_Y)$ be its covering number using balls of radius $\epsilon$. Then the entropy integral is computed from the covering number given by

$$J = J(\delta) := \int_0^1 \sqrt{1 + \log N(\delta\epsilon, B_\delta(h(x)), d_Y)}\, d\epsilon = O(1) \text{ as } \delta \to 0.$$

(R3) There exist constants $\eta > 0, C > 0$, and $\beta > 1$, possibly depending on $x \in (\Omega_X, d_X)$, such that

$$J(\omega, x) - J(h(x), x) \geq C d^{\beta}(\omega, h(x)),$$

for any small neighborhood $d_Y(\omega, h(x)) < \eta$.

The existence of the weak conditional Fréchet means depends on the nature of the space, as well as the metric considered, as discussed in Section 3.1. Assumption (R1) is commonly used to establish the consistency of an M-estimator; see Chapter 3.2 in Van der Vaart and Wellner (2000). In particular, it ensures that weak convergence of the empirical process $\tilde{J}_n$ to the population process $J$, which in turn implies convergence of their minimizers. The conditions on the covering number in Assumption (R2) and curvature in Assumption (R3) arise from empirical process theory and control the behavior of $\tilde{M}_n - M$ near the minimum, which is necessary to obtain rates of convergence. These assumptions are again satisfied for many random objects of interest the common examples of random objects such as distributions, covariance matrices, networks, and so on (see Propositions 1-3 of (Petersen and Müller, 2019)).

**Theorem 3** *Under assumptions (A0)- (A2), (K0), and (R1)- (R2), for any* $x \in (\Omega_X, d_X)$*,*

$$d_Y(\hat{h}(x), h(x)) = o_P(1).$$

**Theorem 4** *Under assumptions (A0)- (A2), (K0), (R1)- (R3), for any* $x \in (\Omega_X, d_X)$*,*

$$d_Y(\hat{h}(x), h(x)) = O_P(\alpha_n),$$

*where* $\alpha_n$ *is as given in* (14)*.*

For most commonly observed random objects $\beta$ in Assumption (R3) is 2, yielding an asymptotic rate of convergence for the M-estimator as $O_P(\alpha_n^{-1})$. With a suitable rate carried from the RKHS regression literature, one can derive the rate of convergence as a function of the sample size $n$. For example, in Li and Song (2017), $\alpha_n \approx n^{-1/4}$, which is improved upon by Sang and Li (2022) as $\alpha_n \approx n^{-1/3}$. This improved rate can be incorporated in the rate of convergence for the weak conditional Fréchet mean to yield an optimal rate of $O_P(n^{-1/3})$.

# 6 Simulation Studies

In this section, we evaluate the numerical performances of the proposed object-on-object regression method under different simulation settings for commonly observed random objects.

In all of the following simulation scenarios, we consider Gaussian radial basis kernel $\kappa_G(y, y') = \exp(-\gamma_X d^2(y, y'))$ as a candidate to construct the underlying RKHS $\mathcal{H}_X$ in the predictor space. We choose the parameters $\gamma_X$ as the fixed quantity

$$\gamma_X = \frac{\rho_Y}{2\sigma_G^2},\ \sigma_G^2 = \binom{n}{2}^{-1} \sum_{i<j} d^2(X_i, X_j),\ \rho_Y = 1.$$

The same choices of tuning parameters were used in Lee et al. (2013); Li and Song (2017); Zhang et al. (2022). The metric $d_X$ and $d_Y$ is chosen appropriately to enhance the interpretability of the results in each of the following scenarios considered.

## 6.1 Scenario 1: Univariate distribution-on-object regression

In the first scenario considered, we have univariate distributional objects as responses coupled with various types of random objects as predictors. Let $(\Omega_Y, d_Y)$ be the metric space of univariate distributions endowed with Wasserstein metric $d_Y = d_W$, as described in (9) Section 3.1. A sample of distributional object response, $Y_1, \ldots, Y_n$ is observed in equivalent forms of CDF, quantile functions, or densities. However, the distributions $Y_1, \ldots, Y_n$ are usually not fully observed in practice and the latent curves need to be recovered from the discrete observations $\{Y_{ij}\}_{j=1}^m$, $i = 1, \ldots, n$, one encounters in reality. For this, we employ nonparametric smoothing with a suitable bandwidth choice implemented by the *CreateDensity()* function in the *frechet* R package (Chen et al., 2020).

The random distributional response $Y$ is generated conditional on $X$ by adding noise to the quantile functions, which are demonstrated in the following simulation settings for various types of predictor objects. Generally, we let $Y = N(\zeta(x), \eta^2(x))$, where the mean and variance of the response distribution are dependent on $X$. To this end, the auxiliary distribution parameters $\mu_Y$ and $\sigma_Y$, given $X$, are independently sampled such that $E(\mu_Y|X = x) = \zeta(x)$ and $E(\sigma_Y^2|X = x) = \eta^2(x)$, and the corresponding distribution is $Y = \mu_Y + \sigma_Y \Phi^{-1}$.

We set $n = 200, 400$, $m = 50, 100$ and generate $n$ samples $(X_i, \{Y_{ij}\}_{j=1}^m)_{i=1}^n$. We use half of them to train the predictors via the proposed object regression method and then evaluate the discrepancy between the estimated and true responses using the rest of the data set by computing the Wasserstein distance metric (9) between the two distributions. The tuning parameter is determined by the methods described in Section 4.3. The experiment is repeated $B = 100$ times, and averages and standard errors (in parentheses) of the prediction error are computed as

$$\text{RMPE} := \frac{1}{B} \sum_{b=1}^{B} d_W(Y_b^{\text{test}}, \hat{Y}_b^{\text{test}}), \tag{15}$$

where $Y_b^{\text{test}}$ and $\hat{Y}_b^{\text{test}}$ are the observed and predicted responses in the test set, respectively, for the $b$-th replicate, $b = 1 \ldots, B$.

**Model-I.1 (Euclidean predictors)**: $\mu_Y|X \sim N((\beta^\intercal X)^2, \nu_1^2)$ and $\sigma_Y|X \sim Gamma((\gamma^\intercal X)^2/\nu_2, \nu_2/(\gamma^\intercal X))$.

**Model-I.2 (Euclidean predictors)**: After sampling the distribution parameters as in the previous setting, the resulting distribution is then "transported" in Wasserstein space via a random transport map $T$, that is uniformly sampled from a family of perturbation/ distortion functions $\{T_k : k \in \pm 1, \pm 2, \}$, where $T_k(x) = x - \sin(kx)/|k|$. The transported distribution is given by $T\#(\mu_Y + \sigma_Y \Phi^{-1})$, where $T\#p$ is a push-forward measure such that $T\#p(A) = p\left(\{x : T(x) \in A\}\right)$, for any measurable function $T : \mathbb{R} \to \mathbb{R}$, distribution $p \in (\Omega_Y, d_W)$, and set $A \subset \mathbb{R}$. We sample the random transport map $T$ uniformly from the collection of maps described above; $p$ denotes a Gaussian distribution with parameters $\zeta(x) = (\beta^{\intercal} X)^2$ and $\eta^2(x) = (\gamma^{\intercal} X)^2$. The distributions thus generated are not Gaussian anymore due to transportation. The Fréchet mean can be shown to remain at $\mu_Y + \sigma_Y \Phi^{-1}$ as before.

For Models I.1 and I.2, the Euclidean vector predictor $X \in \mathbb{R}^p$ is generated as follows: (i) we first generate $U_1, \ldots, U_p$ from the AR(1) model with mean 0 and covariance matrix $\Sigma = (0.5^{|i-j|})_{i,j}$, and then (ii) generate $X_j = 2\Phi(U_j) - 1$, $j = 1, \ldots, p$, where $\Phi$ is the c.d.f. of $N(0, 1)$. We select $\nu_1^2 = 0.1$, $\nu_2 = 0.25$, $\beta = (1, -2, 0, 1)$, and $\gamma = c(0.1, 0.2, 1, 0.3)$ in the above models. The performance of our method, denoted by global nonlinear Fréchet regression (GNLFR), is compared with the globally linear Fréchet regression (GLFR) method by Petersen and Müller (2019) for varying levels of the predictor dimension, sample size, and number of discrete observations for each sample of distributions, namely $p, n$, and $m$, respectively. Table 1 summarizes the results. The prediction error decreases generally corresponding to a lower dimension $p$ of the predictor, a higher sample size $n$, and a denser design (higher $m$) over which the response is sampled. Across the board, our method outperforms the GLFR method in terms of prediction accuracy. Especially, for Setting I.2 the GNLFR method proves significantly better, which is not unexpected given the highly non-linear data-generating mechanism for this setting.

For models I.3-I.5 below, we consider univariate distribution-on-distribution regression.

**Model-I.3 (Univariate distributions as predictors)**: $\mu_Y|X \sim N(\exp(W_2^2(X, \mu_1)) + \exp(W_2^2(X, \mu_2)), \nu_1^2)$ and $\sigma_Y|X = 0.1$.

**Model-I.4 (Univariate distributions as predictors)**: $\mu_Y|X \sim N(\exp(W_2^2(X, \mu_1))\ , \nu_1^2)$ and $\sigma_Y|X = Gamma(W_2^2(X, \mu_2), W_2(X, \mu_2))$.

**Model-I.5 (Univariate distributions as predictors)**: $\mu_Y|X \sim N(\exp(H(X, \mu_1)), 0.2^2); \sigma_Y|X = \exp(H(X, \mu_2))$.

In the above we let $\nu_1^2 = 0.1$, $\mu_1 = Beta(2, 1)$ and $\mu_2 = Beta(2, 3)$ and generate discrete observations from distributional predictors by $\{X_{ij}\}_{j=1}^m \overset{i.i.d.}{\sim} Beta(a_i, b_i)$,

Table 1: Table showing the Monte Carlo mean (standard error) estimation errors as per (15) for the proposed global nonlinear Fréchet regression (GNLFR) and the global linear Fréchet regression by Petersen and Müller (2019) (GLFR), for Euclidean predictors and univariate distributional responses in Scenario I.1-I.2. The lowest number in a row corresponding to each data generating mechanism is highlighted.

| | I.1 (GNLFR) | | I.1 (GLFR) | | I.2 (GNLFR) | | I.2 (GLFR) | |
|---|---|---|---|---|---|---|---|---|
| (p,n)\m | 50 | 100 | 50 | 100 | 50 | 100 | 50 | 100 |
| (4,200) | 0.037<br>(0.012) | **0.024**<br>(0.016) | 0.053<br>(0.021) | 0.038<br>(0.014) | 0.110<br>(0.081) | **0.087**<br>(0.070) | 0.230<br>(0.012) | 0.181<br>(0.011) |
| (10,200) | 0.051<br>(0.019) | **0.042**<br>(0.015) | 0.060<br>(0.017) | 0.049<br>(0.020) | 0.187<br>(0.031) | **0.112**<br>(0.023) | 0.334<br>(0.045) | 0.278<br>(0.031) |
| (20,200) | 0.058<br>(0.018) | **0.051**<br>(0.018) | 0.071<br>(0.020) | 0.065<br>(0.019) | 0.210<br>(0.029) | **0.153**<br>(0.028) | 0.431<br>(0.025) | 0.391<br>(0.022) |
| (4,400) | 0.021<br>(0.009) | **0.013**<br>(0.009) | 0.034<br>(0.010) | 0.021<br>(0.011) | 0.089<br>(0.021) | **0.047**<br>(0.022) | 0.134<br>(0.020) | 0.086<br>(0.021) |
| (10,400) | 0.029<br>(0.010) | **0.023**<br>(0.011) | 0.037<br>(0.009) | 0.024<br>(0.008) | 0.174<br>(0.019) | **0.133**<br>(0.020) | 0.356<br>(0.012) | 0.239<br>(0.014) |
| (20,400) | 0.041<br>(0.013) | **0.033**<br>(0.011) | 0.081<br>(0.015) | 0.043<br>(0.015) | 0.189<br>(0.016) | **0.122**<br>(0.016) | 0.451<br>(0.013) | 0.378<br>(0.015) |

where $a_i \overset{i.i.d.}{\sim} Gamma(2, \text{rate} = 1)$ and $b_i \overset{i.i.d.}{\sim} Gamma(2, \text{rate} = 3)$. $W_2(\cdot, \cdot)$ and $H(\cdot, \cdot)$ denote, respectively, the Wasserstein-2 distance and the Hellinger distance between two univariate distributional objects. The Hellinger distance between two Beta distributions $\mu = Beta(a_1, b_1)$ and $\nu = Beta(a_2, b_2)$ can be represented explicitly as

$$H(\mu, \nu) = 1 - \int \sqrt{f_\mu(t) f_\nu(t)} dt = 1 - \frac{B((a_1 + a_2)/2, (b_1 + b_2)/2)}{\sqrt{B(a_1, b_1) B(a_2, b_2)}},$$

where $B(\alpha, \beta)$ is the $Beta$ function.

Note that by virtue of the Gram matrix of the underlying RKHS kernel $\kappa_X$, the predictor space is now embedded into a Hilbert space, hence finding the weak conditional Fréchet mean reduces to solving a constrained quasi-quadratic optimization problem and projecting back into the solution space.

The performance of our method, denoted by global nonlinear Fréchet regression (GNLFR), is compared with the distribution-on-distribution Wasserstein regression

(WR) proposed by Chen et al. (2021) for varying choices of the sample size and predictor dimension $(n, m)$ (see Table 2). for varying levels of the predictor. The performance of our method is evaluated for varying choices of $(n, m)$ (see Table 2). We observed a decrease in the RMPE as per (15) for all the settings as the sample size $n$ was increased, favorably for the denser design with a higher $m$.

Table 2: Table showing the Monte Carlo mean (standard error) estimation errors as per (15) for univariate distribution-on-distribution regression in Scenario I according to models I.3- I.5. The lowest number in a row corresponding to each data generating mechanism is highlighted.

| | I.3 | | I.4 | | I.5 | |
|---|---|---|---|---|---|---|
| n\m | 50 | 100 | 50 | 100 | 50 | 100 |
| 200 | 0.314<br>(0.121) | **0.268**<br>(0.091) | 0.461<br>(0.110) | **0.381**<br>(0.125) | 0.491<br>(0.110) | **0.407**<br>(0.099) |
| 400 | 0.159<br>(0.092) | **0.134**<br>(0.086) | 0.218<br>(0.160) | **0.172**<br>(0.155) | 0.251<br>(0.181) | **0.177**<br>(0.120) |

We next consider the scenario where $X$ is a two-dimensional random Gaussian distribution in Models I.6-I.7. Similar data generation mechanism was followed in Zhang et al. (2022).

**Model-I.6 (Multivariate distributions as predictors)**: $\mu_Y|X \sim N(\exp(W_2(X, \mu_1)), \nu_1^2)$ and $\sigma_Y|X = 0.1$, with $\mu_1 \sim N((-1, 0)^\intercal, \mathrm{diag}(1, 0.5))$.

**Model-I.7 (Multivariate distributions as predictors)**: $\mu_Y|X \sim N(\exp(W_2(X, \mu_1)), \nu_1^2)$ and $\sigma_Y|X = \tau_1^\intercal \Lambda \tau_2$, with $\mu_1 \sim N((-1, 0)^\intercal, \mathrm{diag}(1, 0.5))$; $\tau_1 = (1/\sqrt{2}, 1/\sqrt{2})^\intercal$, $\tau_2 = (1/\sqrt{2}, -1/\sqrt{2})^\intercal$, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$, where $(\lambda_1, \lambda_2)|X \sim N(W_2(X, \mu_2)(1, 1)^\intercal, 0.25 I_2)$, $\mu_2 \sim N((0, 1)^\intercal, \mathrm{diag}(0.5, 1))$.

When computing $W_2(X, \mu_1)$ and $W_2(X, \mu_2)$, we use the following explicit representations of the Wasserstein distance between two Gaussian distributions:

$$W_2^2(N(m_1, \Sigma_1), N(m_2, \Sigma_2)) = ||m_1 - m_2||^2 + ||\Sigma_1 - \Sigma_2||_F, \tag{16}$$

Table 3 shows a lower RMPE for the less complex setting I.6, while the performance of the method improves for higher $n, m$ as before.

In Model I.8, Hilbertian random functions are taken as predictor objects coupled

Table 3: Table showing the mean and s.e. (in parenthesis) of the prediction errors as per (15), for multivariate distributions as predictors coupled with univariate distributions as responses, as described in Models I.6-I.7 in Scenario I. The lowest number in a row corresponding to each data-generating mechanism is highlighted.

| | I.6 | | I.7 | |
|---|---|---|---|---|
| n\m | 50 | 100 | 50 | 100 |
| 200 | 0.619 (0.110) | **0.534** (0.100) | 0.719 (0.142) | **0.578** (0.131) |
| 400 | 0.467 (0.091) | **0.388** (0.092) | 0.635 (0.110) | **0.541** (0.112) |

with univariate distribution responses, where the distribution of the response varies conditional on the predictor values as before.

**Model-I.8 (Random functions as predictors)**: The predictor trajectories $X$ and associated noisy measurements were generated as follows. Suppose that the simulated process $X$ has the mean function $\mu_X(s) = s + \sin(s)$, with covariance function constructed from two eigenfunctions, $\phi_1(s) = \sqrt{2}\sin(2\pi ks)$ and $\phi_2(s) = \sqrt{2}\cos(2\pi ks)$, $0 \leq s \leq 1$. We chose $\lambda_1 = 1, \lambda_2 = 0.7$ and $\lambda_k = 0$ for $k \geq 3$, as eigenvalues, and the FPC scores $\xi_k$; $(k = 1, 2)$ were generated from $N(0, \lambda_k)$. Using the Kerhunen-Loéve expansion the predictor process is then given by $X(s) = \mu_X(s) + \sum_{k=1}^{\infty} \xi_k \phi_k(s)$. To adequately reflect both a dense design and an irregular/sparse measurement paradigm, we assume that there is a random number $N_i$ of random measurement times for $X_i$ for the $i$-th subject, which are denoted as $S_{i1}, \ldots, S_{iN_i}$ and contaminated with measurement errors $\epsilon_{ij}$, $1 \leq j \leq N_i$, $1 \leq i \leq n$. The errors are assumed to be i.i.d. with $E(\epsilon_{ij}) = 0$ $E[\epsilon_{ij}^2] = \sigma_X^2 = 0.1$, and independent of functional principal component scores $\xi_{ik}$ that satisfy $E[\xi_{ik}] = 0$, $E[\xi_{ik}\xi_{ik'}] = 0$ for $k \neq k'$, and $E[\xi_{ik}^2] = \lambda_k$. Thus, for the $i$-th sample, the predictor measurement with noise is represented as $U_{ij} = \mu_X(S_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(S_{ij}) + \epsilon_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, N_i$. The data generation mechanism above is similar to Yao et al. (2005) and both a sparse and a dense grid of observation are considered with $N_i = 50$ and $N_i \in \{3, \ldots, 5\}$, respectively. Finally, the response as a univariate distribution is constructed as $Y \sim N(\mu_Y, \sigma_Y)$, and the auxiliary parameters conditional on $X(\cdot)$ are generated independently as $\mu_Y|X \sim N((\xi_1, \xi_2)^\intercal \mathrm{diag}(\lambda_1, \lambda_2)(1, -1), \nu_1^2)$ and $\sigma_Y|X = 0.1$.

Again, it is evident from Table 4, that the method yields better prediction error when the sample size and number of discrete observations per sample in the response is high, favorable for the dense design paradigm for the predictor functions.

Table 4: Table showing the average prediction error as per (15) along with the standard error for Hilbertian objects as predictors and univariate distributions as responses, as described in Models I.8 under sparse and dense predictor design. The lowest number in a row is highlighted.

| | I.8 (dense design) | | I.8 (sparse design) | |
|---|---|---|---|---|
| n\m | 50 | 100 | 50 | 100 |
| 200 | 0.334(0.051) | **0.270** (0.049) | 0.483 (0.130) | **0.379** (0.124) |
| 400 | 0.211 (0.031) | **0.176** (0.032) | 0.410 (0.022) | **0.347** (0.022) |

## 6.2 Scenario 2: Multivariate distribution-on-object regression

We now consider the scenario where both $X$ and $Y$ are two-dimensional random Gaussian distributions. The construction of the kernel $\kappa_X$ is done using the sliced 2-Wasserstein distance, which is obtained by computing the average Wasserstein distance of the projected univariate distributions along randomly picked directions. To define formally,

**Definition 5 (Sliced Wasserstein metric)** *let $\mu_1$ and $\mu_2$ be two measures in $\mathcal{P}_p(M)$, the set of Borel probability measures on $(M, \mathcal{B}(M))$ that have finite $p-$th moment and is dominated by the Lebesgue measure on $\mathbb{R}^d$, with $M \subset of \mathbb{R}^d$, $d > 1$. Let $S^{d-1}$ be the unit sphere in $\mathbb{R}^d$. For $\theta \in S^{d-1}$, let $T_\theta : \mathbb{R}^d \to \mathbb{R}$ be the linear transformation $x \mapsto \langle \theta, x \rangle$. Further, let $\mu_1 \circ T_\theta^{-1}$ and $\mu_2 \circ T_\theta^{-1}$ be the push-forward measures by the mapping $T_\theta$. The sliced $p-$Wasserstein distance between $\mu_1$ and $\mu_2$ is then defined by*

$$SW_p(\mu_1, \mu_2) = \left( \int_{S^{d-1}} W_p^p(\mu_1 \circ T_\theta^{-1}, \mu_2 \circ T_\theta^{-1}) d\theta \right)^{\frac{1}{p}}. \tag{17}$$

For $p = 2$, Kolouri et al. (2016) show that the square of sliced Wasserstein distance is conditionally negative definite and hence that the Gaussian RBF kernel defined as $\kappa_X(x, x') = \exp(-\gamma_X SW_2^2(x, x'))$ is a positive definite kernel.

We generate discrete observations for the predictor distributions $X_i, i = 1, \ldots, n$ given by $\{X_{ij}\}_{j=1}^m \overset{i.i.d.}{\sim} N(a_i(1,1)^T, b_i I_2)$, where $a_i \overset{i.i.d.}{\sim} N(0.5, 0.5^2)$ and $b_i \overset{i.i.d.}{\sim} Beta(2,3)$.

For computing the Gram matrix associated with the multivariate predictor distribution supported on $M \subset \mathbb{R}^d$, $d > 1$ the sliced Wasserstein distance is estimated using a Monte Carlo method as

$$SW_2(\mu_{X_i}, \mu_{X_k}) \approx \left( \frac{1}{L} \sum_{l=1}^{L} W_2^2(\mu_{X_i} \circ T_\theta^{-1}, \mu_{X_k} \circ T_\theta^{-1}) \right)^{\frac{1}{2}},$$

where $\mu_{X_i} = \frac{1}{m} \sum_{j=1}^{m} \delta_{X_{ij}}$ is the empirical measure for the $i-$th sample, $i = 1, \ldots, n$, $\{\theta_l\}_{l=1}^{L}$ are i.i.d. samples drawn from the uniform distribution on $S^{d-1} \subset \mathbb{R}^d$. The approximation error depends on the number of Monte Carlo samples $L$. In our simulation settings, we set $L = 50$.

The random responses $Y = N(\mu_Y, \Sigma_Y)$, where $\mu_Y \in \mathbb{R}^2$ and $\Sigma_Y \in \mathbb{R}^{2\times 2}$ are then generated according to the following models.

**Model-II.1 (Multivariate distributions as predictors)**: $\mu_Y | X \sim N(W_2(X, \mu_1)(1, 1)^\intercal, I_2)$ and $\Sigma_Y | X = \mathrm{diag}(1, 1)$.

**Model-II.2 (Multivariate distributions as predictors)**: $\mu_Y | X \sim N(W_2(X, \mu_1)(1, 1)^\intercal, I_2)$ and $\Sigma_Y | X = \Gamma\Lambda\Gamma^\intercal$, where $\Gamma = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$ with $(\lambda_1, \lambda_2)|X \overset{i.i.d.}{\sim} tGamma(W_2^2(X, \mu_2), W_2(X, \mu_2), (0.2, 2))$, where $\mu_1$ and $\mu_2$ are two fixed measures defined by $\mu_1 = N((-1, 0)^\intercal, \mathrm{diag}(1, 0.5))$ and $\mu_2 = N((0, 1)^\intercal, \mathrm{diag}(0.5, 1))$, and $tGamma(\alpha, \beta, (r_1, r_2))$ is the truncated gamma distribution on range $(r_1, r_2)$ with shape parameter $\alpha$ and rate parameter $\beta$. The Wasserstein distance between the bivariate Gaussian distributions is computed as per (16).

Since the dimension $d$ of the random probability measures that we study here is more than 1, one does not have an analytic form for the barycenter and the optimization algorithms to obtain it are complex, in contrast to the case $d = 1$, where the quantile representation of Wasserstein distance leads to an explicit solution via the $L_2$ mean of the quantile functions. The computation of Wasserstein barycenters in multidimensional Euclidean space has been intensively studied (e.g., Rabin et al. (2012); Álvarez-Esteban et al. (2016); Dvurechenskii et al. (2018); Peyré et al. (2019), and one of the most popular methods utilize the Sinkhorn divergence (Cuturi, 2013), which is an entropy-regularized version of the Wasserstein distance that allows for computationally efficient solutions of the barycenter problem, however at the cost of introducing a bias, as is common for regularized estimation. Due to the gain in efficiency, we adopt this approach in our implementations using the R package *WSGeometry* (Heinemann and Bonneel, 2021).

Using the same choices for $n$, $m$, and the tuning parameters, we again split the data into a training and a test set, and use the training set to implement the proposed object regression method at the output predictor points to predict the response in the testing set. The whole process is repeated $B = 100$ times and the prediction error computed between the observed and predicted bi-variate distributional responses in the test set using the average Sliced Wasserstein distance between them, as per (17). The mean and standard error of this root mean prediction error is shown in Table 5, where a similar pattern of decreased RMPE for a combination of higher sample size and denser observation grid for the paired sample of distribution is noted.

Table 5: Table showing the Monte Carlo mean (standard error) prediction errors for Scenario II. The lowest number in a row is highlighted across different model settings.

| | II.1 | | II.2 | |
|---|---|---|---|---|
| n\m | 50 | 100 | 50 | 100 |
| 200 | 0.620 (0.134) | **0.442** (0.130) | 0.811 (0.200) | **0.693** (0.177) |
| 400 | 0.319 (0.094) | **0.178** (0.092) | 0.543 (0.160) | **0.329** (0.152) |

## 6.3 Scenario 3: SPD matrix object-on-object regression

A common type of random object encountered in brain imaging studies is functional connectivity correlation matrices, which are positive semi-definite symmetric matrices. Let $(\Omega_Y, d_F)$ be the space of $r \times r$ symmetric positive definite (SPD) matrices endowed with Frobenius distance $d_F(Y_1, Y_2) = ||Y_1 - Y_2||_F$ as defined in (10) in Section 3.1. Two simulation scenarios are considered as follows.

**Model-III.1 (Euclidean predictors)**: The real-valued predictors $X_i$ are independently sampled from a $Beta(1/2, 2)$, while the SPD matrix responses $Y_i$ conditional on $X_i$ are generated according to the model $Y_i = \tilde{Y}_i\tilde{Y}_i^T$, with $\tilde{Y}_i|X_i = \mu(X_i) + [\Sigma(X_i)]^{-1/2}Z_i$, where for a fixed dimension $r$, the mean vector $\mu(x)$ has components $\mu_j(x) = b_j - 2(x - c_j)^2$, $j = 1, \ldots, r$. Here $b_j \sim U(2, 4)$ and $c_j \sim U(0, 1)$, and $Z_i$ are sampled independently of $X_i$ as a standard $r-$dimensional Gaussian random vector. the covariance $\Sigma(x)$ is formed by generating a $r \times r$ matrix $A$ with independent $N(0, 0.5)$ random variables in each entry, then computing $S = 0.5(A + A^T)$. A second $r \times r$ matrix $V$ is generated with elements drawn independently as $U(0, 0.5)$, from which $\theta = 0.5(V + V^T)$ is computed. Finally, with *Exp* denoting matrix exponentiation and $\odot$ the Hadamard product, we form $\Sigma(x) = (x + 2x^3)Exp[S \odot \sin(2\pi\theta(x + 0.1))]$.

**Model-III.2 (SPD matrix objects as predictors)**: The predictors are now themselves SPD matrices. This is generated as the covariance matrix computed from a $p$-variate Gaussian random vector with independent components each with mean 0 and variance 1 for each sample. The predictors are projected down on a desired direction vector $\beta$ whose each component $\beta_j \sim U(0, 1)$, $j = 1, \ldots, p$ to compute $\tilde{X}_i = X_i\beta$. Here, we choose $p = 5$. Now the response matrices are generated as before in Model III.2 conditional on $\tilde{X}_i$.

In order to apply the proposed method, again the Gaussian RBF kernel given by $\kappa_X(x, x') = \exp(-\gamma_X d_F^2(x, x')$ is taken to compute the Gram matrix in the predictor space, with the tuning parameter chosen as before. From a sample $(X_i, Y_i)_{i=1}^n$ the minimization in (12) can be reformulated by setting $\hat{h}(x) = \frac{1}{n}\sum_{i=1}^n w_{in}(x)Y_i$ and computing the correlation matrix which is nearest to the matrix $\hat{h}(x)$, which is implemented by the alternating projections algorithm via the *nearPD()* function in the *Matrix* R package.

We compare performances of the proposed method for a combination of sample size and the dimension of the response matrices given by $n$ and $r$, respectively, by computing the Frobenius distance between the true and the predicted SPD matrix responses in the test set, using the model fit on the training set, as described before. The first two columns of Table 6 display the average prediction error across 100 replications of the above process. Our method fares better for increased sample size, while the dimension of the response SPD matrices is lower in both simulation scenarios.

Table 6: Table showing the Monte Carlo mean (standard error) estimation errors for Scenarios III and IV. The lowest number in a row is highlighted across different model settings.

| | III.1 | | III.2 | | IV.1 | |
|---|---|---|---|---|---|---|
| n\r | 5 | 20 | 5 | 20 | 5 | 20 |
| 200 | **0.119**<br>(0.041) | 0.275<br>(0.040) | **0.226**<br>(0.130) | 0.786<br>(0.110) | **0.161**<br>(0.011) | 0.235<br>(0.031) |
| 400 | **0.048**<br>(0.037) | 136<br>(0.035) | **0.127**<br>(0.110) | 0.502<br>(0.097) | **0.079**<br>(0.012) | 0.145<br>(0.029) |

## 6.4 Scenario 4: Network object-on-object regression

**Model-IV.1 (Euclidean predictors)**: Let $G = (V, E)$ be a simple (with no self-loops), weighted, undirected network with a set of nodes $V = \{v_1, \ldots, v_r\}$ and a set of edge weights $E = \{w_{ij} : w_{ij} \geq 0,\ i, j = 1, \ldots, r\}$, where $w_{ij} = 0$ indicates $v_i$ and $v_j$ are not connected and $w_{ij} > 0$ otherwise, with $w_{ij} < M$ for some $M > 0$. A network can be uniquely represented by its graph Laplacian $L = (l_{ij})$, where $l_{ij} = -w_{ij}$ if $i \neq j$ and $l_{ij} = \sum_{k \neq i} w_{ik}$ if $i = j$, for $i, j = 1, \ldots, r$. The space of graph Laplacians is given by $\mathcal{L}_r = \{L = (l_{ij}) : L = L^\intercal,\ L1_r = 0_r,\ -W \leq l_{ij} \leq 0 \text{ for some W } \geq 0 \text{ and } i \neq j\}$, where $1_r$ and $0_r$ are the $r$-vectors of ones and zeroes, respectively. Note that $\mathcal{L}_r$ is not a linear space, but a bounded, closed, and convex subset in $\mathbb{R}^{r^2}$ of dimension $r(r-1)/2$. Owing to the fact that $x^\intercal L x \geq 0$ for all $x \in \mathbb{R}^r$ and $L \in \mathcal{L}_r$, it can be seen as a metric space of positive-semidefinite matrix objects, equipped with a suitable choice of metric such as the Frobenius or power metric.

To assess the performance of our proposed methods, we consider the space $(\mathcal{L}_r, d_F)$, where $d_F$ is the Frobenius metric as per (10). The data generation mechanism is as follows. Denote the half vectorization excluding the diagonal of a symmetric and centered matrix by $vech$, with inverse operation $vech^{-1}$. By the symmetry and centrality, every graph Laplacian $L$ is fully known by its upper (or lower) triangular part, which can then be vectorized into $vech(L)$, a vector of length $d = r(r-1)/2$. We construct the conditional distributions $F_{L|X}$ by assigning an independent beta distribution to each element of $vech(L)$. Specifically, a random sample $(\beta_1, \ldots, \beta_d)^\intercal$ is generated using beta distributions whose parameters depend on the scalar predictor $X$ and vary under different simulation scenarios. The random response $L$ is then generated conditional on $X$ through an inverse half vectorization $vech^{-1}$ applied to $(\beta_1, \ldots, \beta_d)^\intercal$. The the true regression function $m(x)$ is defined as $m(x) = vech^{-1}(-x, \ldots, -x)$, $L = vech^{-1}(\beta_1, \ldots, \beta_d)^\intercal$, where $\beta_j \overset{i.i.d.}{\sim} Beta(X, 1-X)$. To ensure that the random response $L$ generated in simulations resides in $\mathcal{L}_r$, the off-diagonal entries $-\beta_j$ $j = 1, \ldots, d$, need to be nonpositive and bounded below. Thus we choose $\beta_j \overset{i.i.d.}{\sim} Beta(X, 1-X)$. The scalar predictor $X_i$ are randomly sampled from a $Unif(0, 1)$ distribution to obtain the samples of pairs $(X_i, L_i)$, $i = 1, \ldots, n$, setting $r = 5, 20$, and following the above procedure. The prediction error w.r.t the Frobenius metric is shown in the rightmost column of Table 6. The method performs better for higher $n$ and lower $r$.

# 7 Data Analysis

In this application, we explore the relationship between the distribution of age-at-death and that of the mother's age at birth at a country level. Going beyond summary statistics such as mortality or fertility rate, viewing the entire distributions as samples of data is more informative and insightful to understanding the nature of human longevity and its dependence on relevant predictors. The data is obtained from the UN World Population Prospects 2019 Databases (`https://population.un.org`). For this analysis, we focus on $n = 194$ countries over the period of time $2015 - 2020$. The mortality data is available in the form of life tables over the age interval $[0, 110]$ (all in years) while the number of births is categorized by the mother's age every five years over the age bracket $[15, 50]$. We used bin widths equal to 5 years to construct the histograms for the mortality and fertility distributions, respectively, and proceeded to obtain the smooth densities by applying local linear regression using the *frechet* package at the country level. The domains of the age-at-death and mother's age-at-birth densities are $[0, 110]$ and $[15, 50]$ years, respectively. The densities are assumed to lie in the space of univariate distributions equipped with the Wasserstein metric $(\Omega_Y, d_W)$ in (9). Figure 1 shows the sample of densities observed.

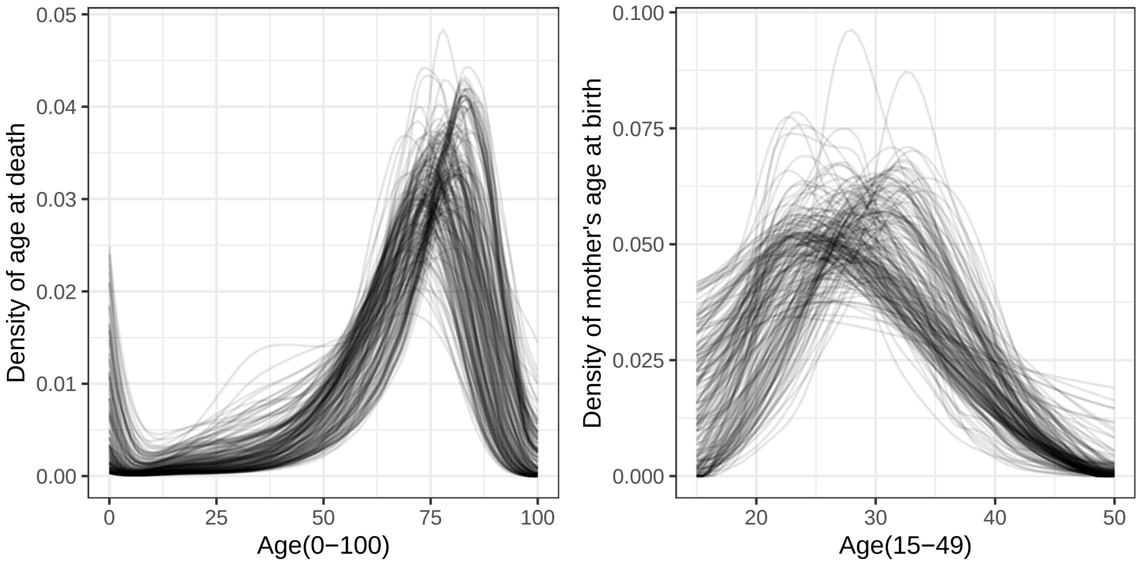


Figure 1: Visualization of distributional objects represented as densities of age at death and mother's age at birth for a sample of 194 countries.

We applied the proposed object-on-object regression method with age-at-death densities as responses and mother's age-at-birth densities as predictors to compare the evolution of mortality distributions among different countries. We show the leave-one-out prediction results together with the observed distributional predictors and responses in Figure 2 for a select few countries, which showcases different patterns of mortality change over changes in the predictor distribution. The Wasserstein distance between the observed and predicted distributions is also shown. Specifically, we selected the countries Bangladesh, Argentina, the USA, Japan, the UK, and Norway, ordered by the lowest to the highest value of the mode of the mother's age-at-death densities. Both the observed and predicted age-at-death densities across the panels from left to right are seen to be more right-shifted, indicating increased longevity corresponding to a higher age at birth for the mother. Further, for Japan, Norway, and the USA, the rightward mortality shift is seen to be more expressed than suggested by the prediction, indicating that longevity extension is more than anticipated, while the mortality distribution for the UK seems to shift to the right at a slower pace than predicted, leading to a relatively larger WD with a value of 0.8 between the observed and predicted response. In contrast, the regression fit for Argentina and Bangladesh are quite accurate.

The effect of the mother's age-at-birth is elicited in Figure 3a, where the model is fitted for varying levels of the mode of the predictor distribution. The fitted densities are color coded such that blue to red indicates smaller to larger values of the mode of the age-at-birth densities. We find that lower age-at-birth of the mother is associated with left-shifted age-at-death distributions in general, while modes at higher age-at-birth correspond to a shift of the mode of the age-at-death toward the right. Child mortality has an association with both low and high values of age-at-birth for the mother, which concurs with the observations made earlier.

The fit of the model is further demonstrated by computing the estimation error by virtue of the residual map for the $i$-th subject, $T_i : \Omega_Y \to \Omega_Y$, defined as the optimal transport map $T_i = \nu_i \# \hat{\nu}_i$, that pushes forward the observed response $\nu_i$ to the fitted value $\hat{\nu}_i$. Using the theory of optimal transport for univariate distributions (Villani et al., 2009), this map can be explicitly computed as $T_i = Q_{\hat{\nu}_i} \circ F_{\nu_i}$, where $Q_{\hat{\nu}_i}$ and $F_{\nu_i}$ are, respectively, the quantile function and the CDF of the distributions $\hat{\nu}_i$ and $\nu_i$. Using these residual maps one can obtain an analog of the "residual plot" in the classical regression case, compared to the identity map. Looking at the deviation

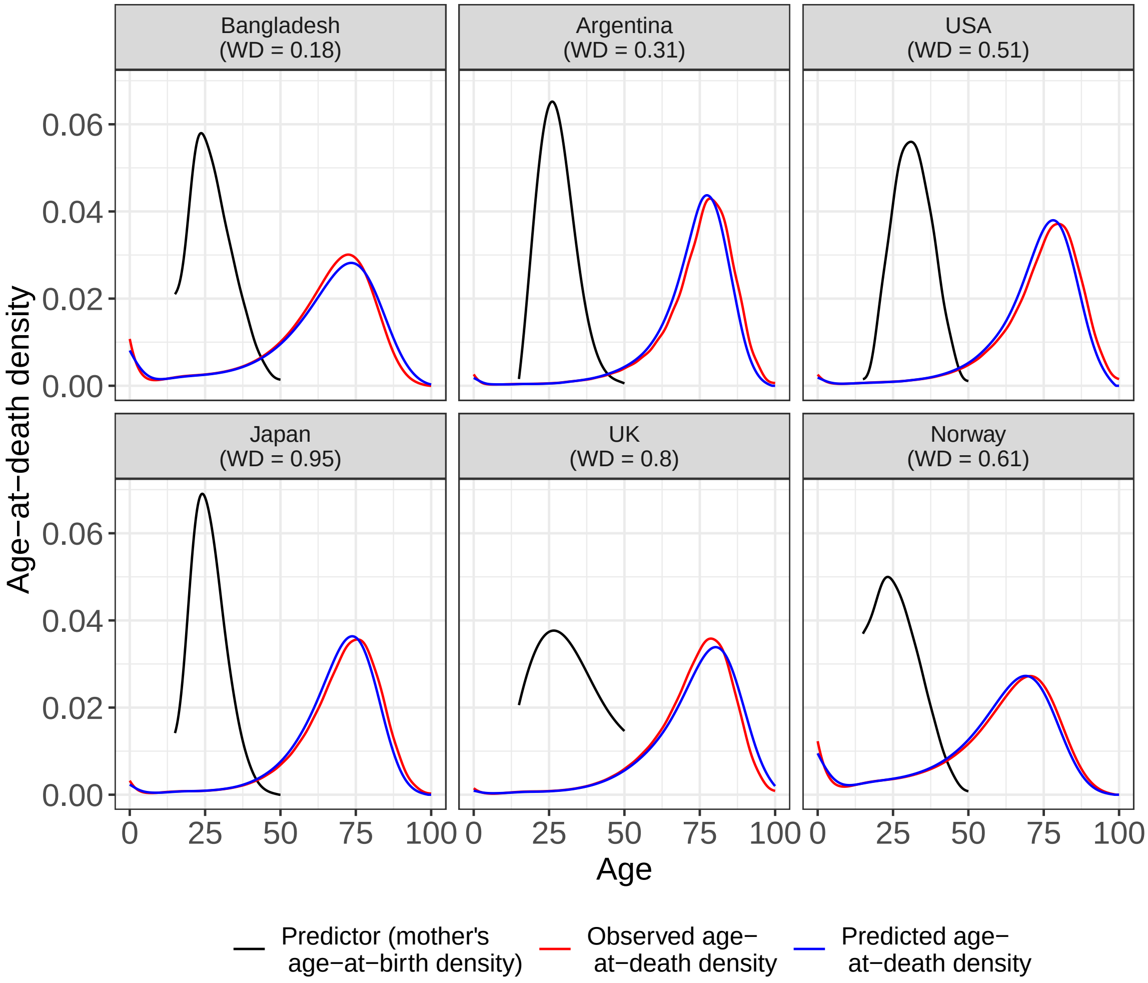
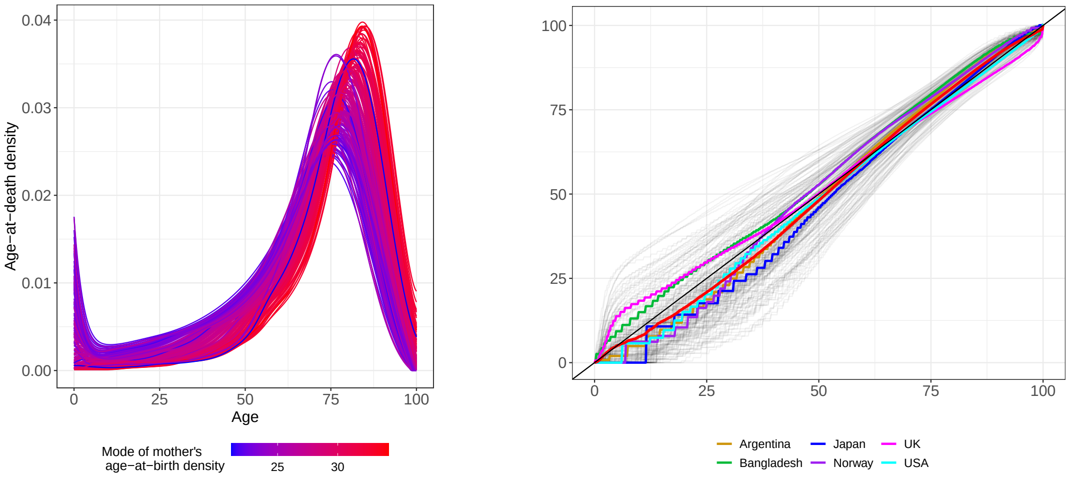


Figure 2: Visualization of distributional objects represented as densities of age at death and mother's age at birth for a sample of 194 countries.

from the identity map one can see in which parts the support of the distributions, the model provides a good fit, and where less so and the departure from the identity can serve as a diagnostic tool for the validity of the model. Note that, contrary to classical regression, where the residuals add up to zero by construction, the residual maps are not constrained to have a mean equal to the identity.

The residual maps computed for each of the 194 countries are plotted in Figure 3b. One can see that the pointwise variability is much more prominent for younger ages and decreases for progressively older ages, indicating many other plausible factors affecting mortality at younger ages. The identity map is overlaid in black. The mean



(a) The changes in the density of the age-at-death distribution as the mode of the distribution of the mother's age-at-birth ranges from low (blue) to high (red) are displayed.

(b) Residual maps corresponding to $n = 197$ countries are plotted in gray, with specific countries highlighted. The identity map and the average residual map are overlaid in black and red, respectively.

of residuals plotted in red lies very close to the identity map, which provides evidence in support of the validity of our model. The residual maps of the specific countries considered in Figure 2 are highlighted. Similar patterns of right-shifted distributions, especially near the age-at-death $[15, 40]$ years are observed for the highlighted countries. For example, while the evolution of the mortality distributions for Japan and the USA can be viewed as mainly a rightward shift over calendar years, this is not the case for the UK, where compared with the fitted response, the actual rightward shift of the mortality distribution seems to be accelerated for those above age 65, and decelerated for those below age 65.

To evaluate the out-of-sample prediction performance of the method, we randomly split the dataset into a training set and a test set, and use the fits obtained from the training set to predict the responses to the test set using only the predictors present in the test set. As a measure of the efficacy of the fitted model, we compute the root mean squared prediction error (RMPE) as the Wasserstein discrepancy between the observed and the predicted distributions in the test set. We repeat the process 100 times to obtain the average RMPE, which comes out low $(0.693$ with a standard error of $0.151)$, supporting the efficacy of the model.

# 8 Discussion

In this paper, we have proposed a nonlinear global object-on-object regression method based on the intrinsic geometry of the metric space where the responses reside coupled with suitable linear operators defined via the reproducing kernel Hilbert space on the predictor space. This contribution is one of the first to model the regression relationship between metric-valued objects, beyond scalar-or-vector-valued predictors. Further, the lack of linearity in an abstract metric space can result in a significant difference between conditional and globally linear Fréchet means proposed by Petersen and Müller (2019), leading to questions about the validity of such globally linear models. To address this, we introduce a novel method extending global linear regression to a general global non-linear object regression. We employ generalized weak conditional Fréchet moments as a way to link random object data analysis to non-linear global RKHS regression models, allowing for arbitrary non-linear functions beyond linear or polynomial regression.

The concept of weak Fréchet moments can be easily extended to Fréchet median or as a minimizer of Huber loss, by substituting $E[d_Y^2(Y, \cdot) ¦ X]$ by $E[\rho_Y(Y, \cdot) ¦ X]$, for any appropriate convex loss function $\rho_Y$ in the metric space $(\Omega_Y, d_Y)$, depending on the context and interpretation of the problem. This calls for future research. The selection of a suitable metric is also an open problem.

Further, the rate of convergence of the proposed estimator is derived as $\approx n^{-1/4}$, which entails from the work of Li and Song (2017). This rate can be further improved using a suitable rate carried out from the RKHS regression literature.

# References

Afsari, B. (2011). Riemannian $L^p$ center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673.

Álvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2016). A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762.

Bhattacharjee, S. and Müller, H.-G. (2021). Single index fr\'echet regression. *arXiv preprint arXiv:2108.05437*.

Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767.

Chen, Y., Gajardo, A., Fan, J., Zhong, Q., Dubey, P., Han, K., Bhattacharjee, S., and Müller, H. (2020). frechet: statistical analysis for random objects and non-euclidean data. *R package version 0.2. 0.*

Chen, Y., Lin, Z., and Müller, H.-G. (2021). Wasserstein regression. *Journal of the American Statistical Association*, pages 1–14.

Chen, Z., Bao, Y., Li, H., and Spencer Jr, B. F. (2019). Lqd-rkhs-based distribution-to-distribution regression methodology for restoring the probability distributions of missing shm data. *Mechanical Systems and Signal Processing*, 121:655–674.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Delicado, P. and Vieu, P. (2017). Choosing the most relevant level sets for depicting a sample of densities. *Computational Statistics*, 32(3):1083–1113.

Di Marzio, M., Panzera, A., and Taylor, C. C. (2014). Nonparametric regression for spherical data. *Journal of the American Statistical Association*, 109(506):748–763.

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Annals of Applied Statistics*, 3:1102–1123.

Dryden, I. L., Pennec, X., and Peyrat, J.-M. (2010). Power euclidean metrics for covariance matrices with application to diffusion tensor imaging. *arXiv preprint arXiv:1009.3045*.

Dubey, P. and Müller, H.-G. (2019). Fréchet analysis of variance for random objects. *Biometrika*, 106(4):803–821.

Dvurechenskii, P., Dvinskikh, D., Gasnikov, A., Uribe, C., and Nedich, A. (2018). Decentralize and randomize: Faster algorithm for wasserstein barycenters. *Advances in Neural Information Processing Systems*, 31.

Fréchet, M. R. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri Poincaré*, 10(4):215–310.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(2).

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99.

Ghodrati, L. and Panaretos, V. M. (2022). Distribution-on-distribution regression via optimal transport maps. *Biometrika*, 109(4):957–974.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Hein, M. (2009). Robust nonparametric regression with metric-space valued output. *Advances in neural information processing systems*, 22.

Heinemann, F. and Bonneel, N. (2021). Wsgeometry: compute wasserstein barycenters, geodesics, pca and distances. *R package version 0.1. 0*.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Kolouri, S., Zou, Y., and Rohde, G. K. (2016). Sliced wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267.

Le Gouic, T. and Loubes, J.-M. (2017). Existence and consistency of wasserstein barycenters. *Probability Theory and Related Fields*, 168(3):901–917.

Lee, K.-Y., Li, B., and Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation.

Lee, K.-Y., Li, B., and Zhao, H. (2016). Variable selection via additive conditional independence. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 1037–1055.

Li, B. (2018). *Sufficient dimension reduction: Methods and applications with R*. CRC Press.

Li, B. and Song, J. (2017). Nonlinear sufficient dimension reduction for functional data.

Li, B. and Song, J. (2022). Dimension reduction for functional data based on weak conditional moments. *The Annals of Statistics*, 50(1):107–128.

Lin, Z. (2019). Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370.

Marron, J. S. and Alonso, A. M. (2014). Overview of object oriented data analysis. *Biometrical Journal*, 56(5):732–753.

Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics*, 47(2):691–719.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2012). Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer.

Sang, P. and Li, B. (2022). Nonlinear function-on-function regression by rkhs. *arXiv preprint arXiv:2207.08211*.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561.

Su, Z., Li, B., and Cook, D. (2023). Envelope model for function-on-function linear regression. *Journal of Computational and Graphical Statistics*, pages 1–12.

Van der Vaart, A. and Wellner, J. (2000). *Weak Convergence and Empirical Processes: with Applications to Statistics (Springer Series in Statistics)*. Springer, corrected edition.

Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer.

Weidmann, J. (2012). *Linear operators in Hilbert spaces*, volume 68. Springer Science & Business Media.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data.

Zhang, Q., Li, B., and Xue, L. (2022). Nonlinear sufficient dimension reduction for distribution-on-distribution regression. *arXiv preprint arXiv:2207.04613*.

Zhang, Q., Xue, L., and Li, B. (2021). Dimension reduction and data visualization for fréchet regression. *arXiv preprint arXiv:2110.00467*.